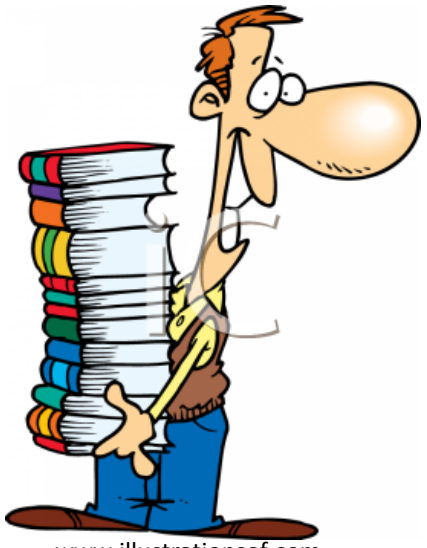
Analitička statistika
Testiranje hipoteze



www.illustrationsof.com

Dijelovi istraživanja

- Istraživačko pitanje
- Značenje
- Ustroj (*design*) - tip istraživanja
- Ispitanici
- Varijable
- Statistička obrada podataka – testiranje hipoteze

© Lucasfilm

Procjena na temelju uzorka

- Pogrešno zaključivanje o uzročnoj povezanosti može nastati zbog:
  - Slučajne pogreške (engl. *random error*)
  - Sustavne pogreške (engl. s*ystematic error*)
  - Zabune (engl. *confounding*)

Procjena

- Slučajna pogreška → niska preciznost

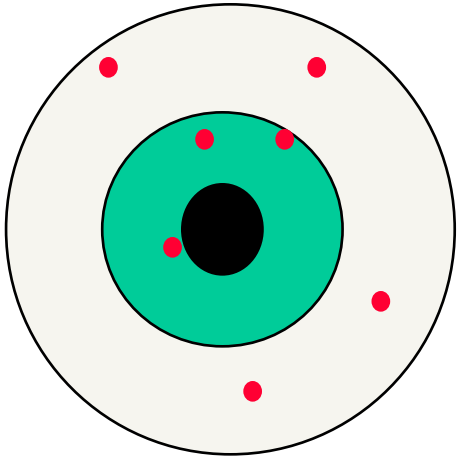
- Sustavna pogreška → niska točnost (validnost)

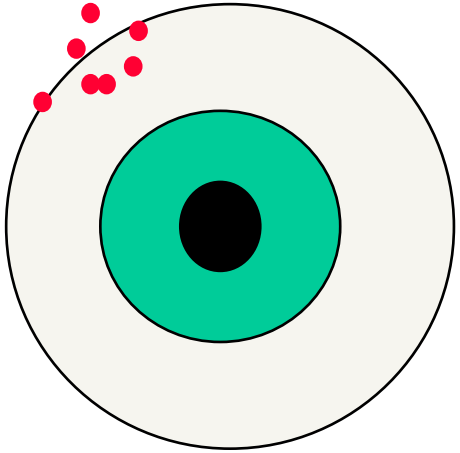
Procjena - osnovni pojmovi

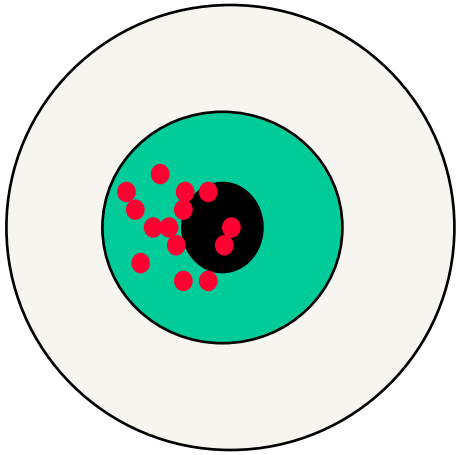
Niska preciznost

Dobra preciznost, ali niska točnost

Dobra preciznost, dobra točnost







Slučajna pogrješka

Sustavna pogrješka

Procjena - standardna pogreška

- Procjenjuje preciznost rezultata
- **Ne** procjenjuje točnost podataka!!!
- SEM ili SE(*p*):

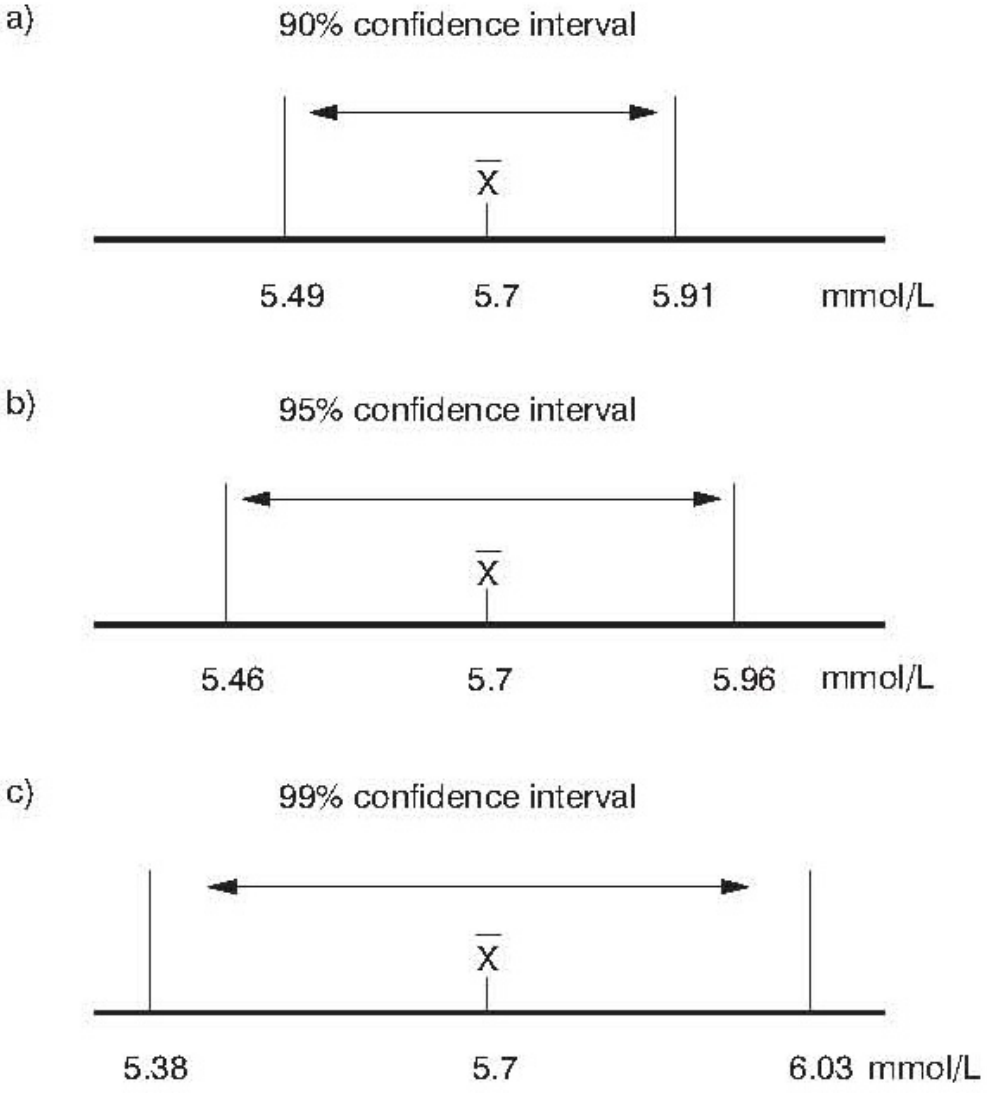
$$\mathrm{SEM} = \mathrm{SD}/\sqrt{n}$$

$$\mathrm{SE}(p) = \sqrt{(p(1-p)/n)}$$

Procjena - raspon pouzdanosti

- Objedinjuje i preciznost i točnost procjene
- Raspon vrijednosti unutar kojeg s određenom sigurnošću možemo reći da se nalazi prava vrijednost mjerenog svojstva u populaciji
- Primjer 1: arit. sredina iznosi 152, a 95% CI 131 – 173
- Primjer 2: arit. sredina iznosi 152, a 95% CI 140 – 164

Procjena - raspon pouzdanosti



Procjena - raspon pouzdanosti

Za izračun treba znati:
Srednju vrijednost, X
Standardnu devijaciju, σ
Veličinu uzorka, n

(tj. standardnu pogrješku)

$$\text{Lower endpoint} = \bar{X} - 1.96\frac{\sigma}{\sqrt{n}},$$

$$\text{Upper endpoint} = \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}.$$

Confidence level	Z value
90%	1.65
95%	1.96
99%	2.58
99,9%	3.291

Podjela statistike…

- Statistika
  - Deskriptivna
  - Analitička/inferencijalna
    - Parametrijska – za normalnu raspodjelu
    - Neparametrijska – za raspodjelu koja odstupa od normalne

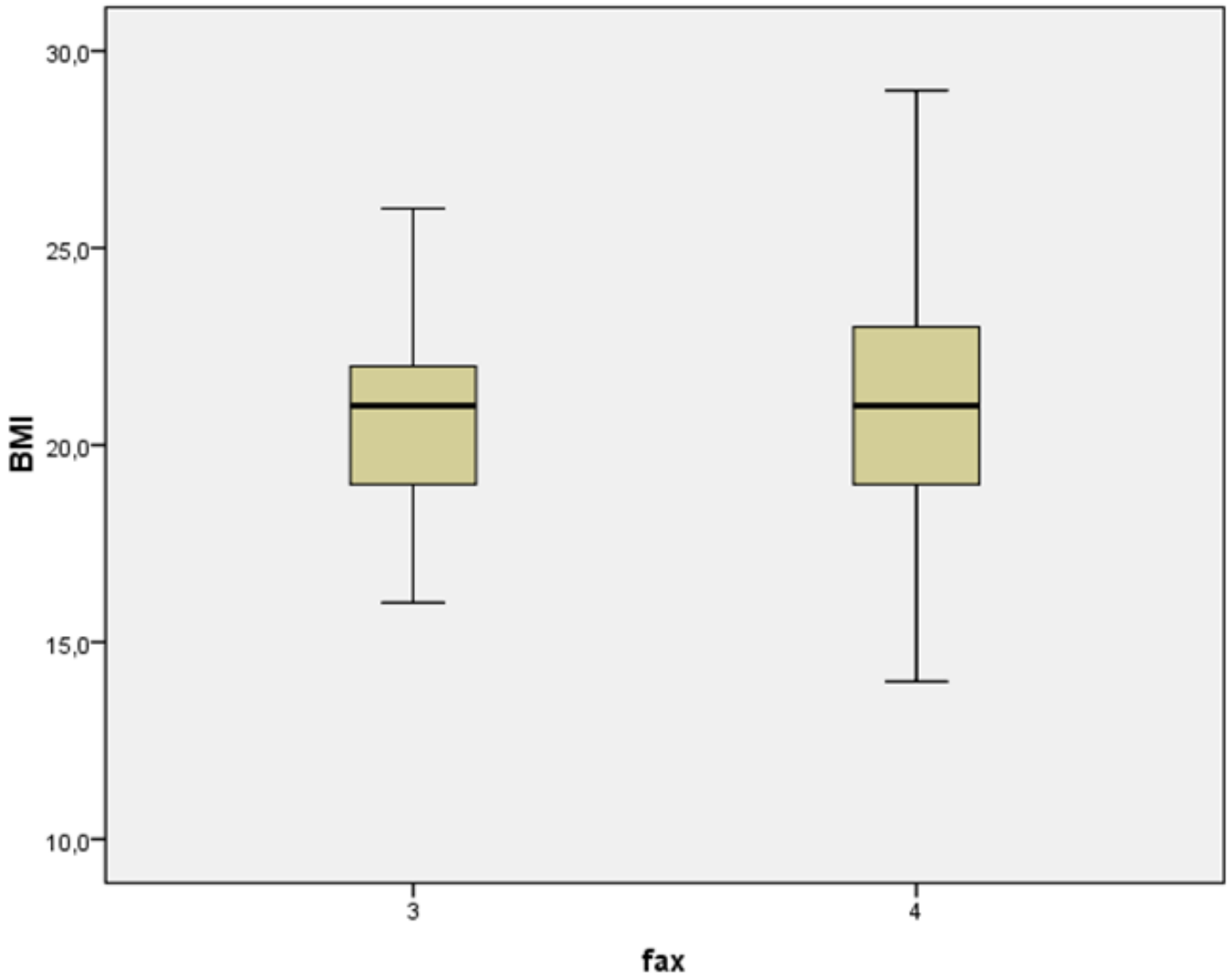
Deskriptivna statistika

- Prikaz mjera središnje vrijednosti
- Prikaz mjera varijabilnosti podataka (rasap)
- **UVIJEK ZAJEDNO!**

- Normalna raspodjela: srednja vrijednost±standardna devijacija
- Raspodjela podataka koja odstupa od normalne: medijan i (1) raspon, (2) najmanja i najveća vrijednost i (3) interkvartilni raspon

Medijan i mjere varijabilnosti

- Medijan (raspon) max-min
  - 56,0 (75,0)
- Medijan (raspon) min i max
  - 56,0 (18,0-93,0)
- Medijan (interkvartilni raspon; 75’-25’)
  - 56,0 (24,0)


30,0
25,0
20,0
15,0
10,0
BMI
3
4
fax

Testiranje hipoteze

- Što je hipoteza?
- H_0 – ništična (nul-hipoteza) = negacijska
- H_1 – alternativna = afirmacijska

- Npr. istraživačko pitanje: smanjuje li uzimanje vitamina C rizik za prehladu?

- H_0: uzimanje vitamina C ne smanjuje rizik za prehladu
- H_1: uzimanje vitamina C smanjuje rizik za prehladu

Testiranje hipoteze – pravilan redoslijed?

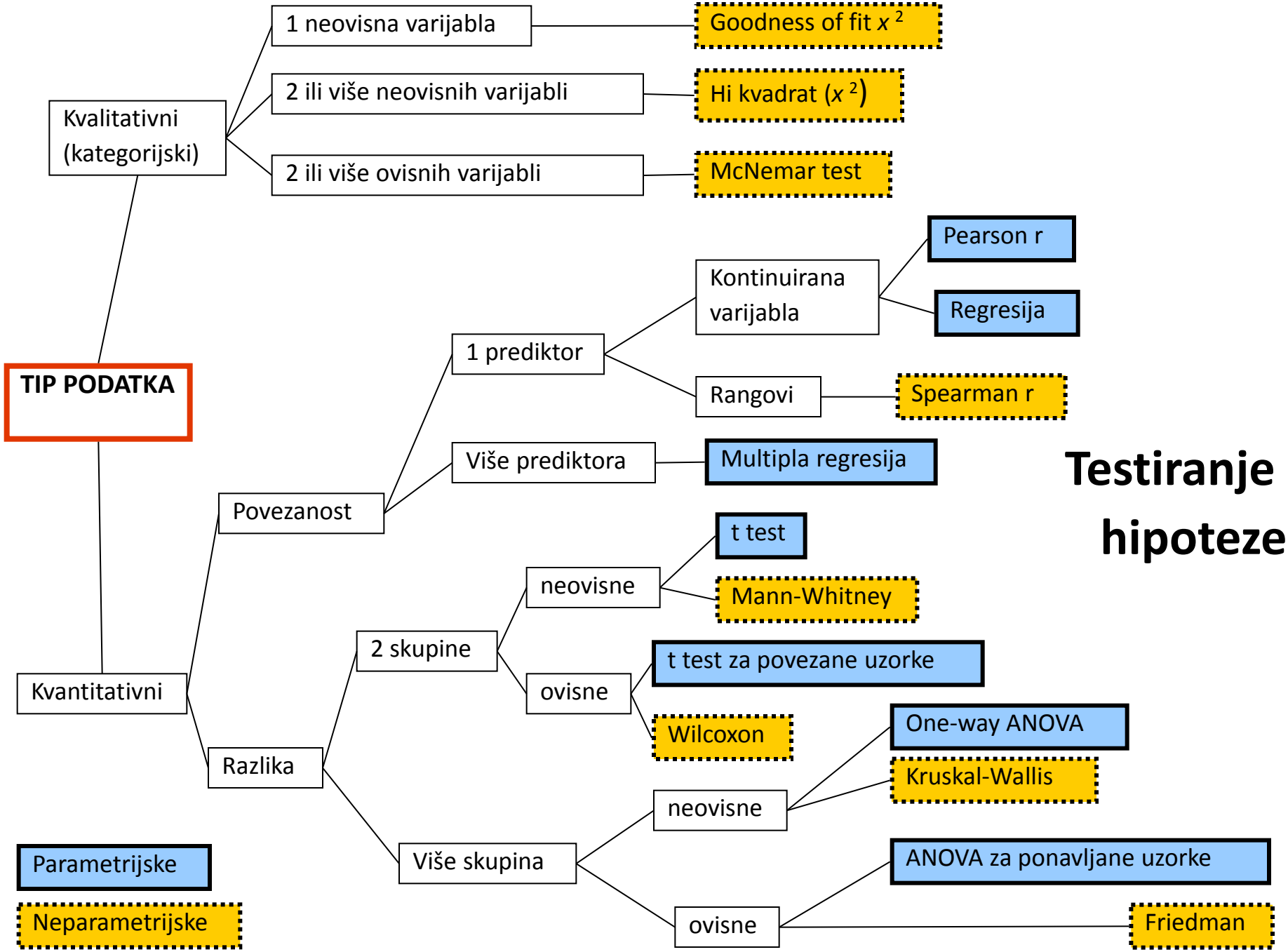
A. Tumačenje P-vrijednosti
B. Statistički izračun
C. Postavljanje ništične i alternativne hipoteze
D. Prikupljanje odgovarajućih podataka
E. Očitavanje P-vrijednosti iz odgovarajuće krivulje raspodjele vjerojatnosti

Statistički izračun

- Za proveden statistički test dobijemo:

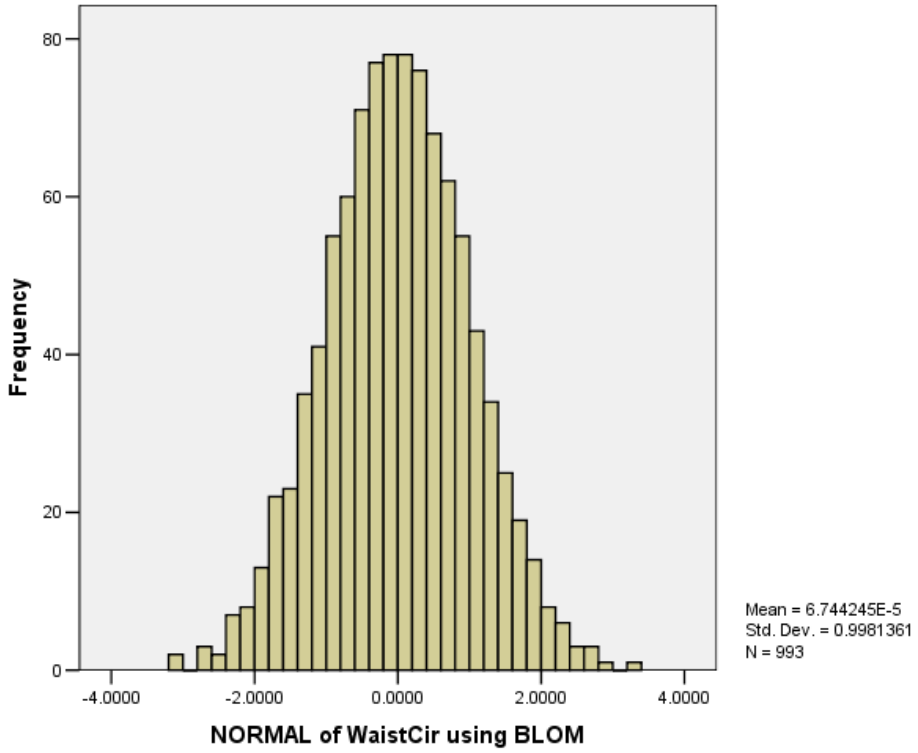
1) rezultat statističkog testa (*test statistic*)
2) P vrijednost ili 95% CI

Npr.: $\chi^2 = 20,3$; $P<0,001$

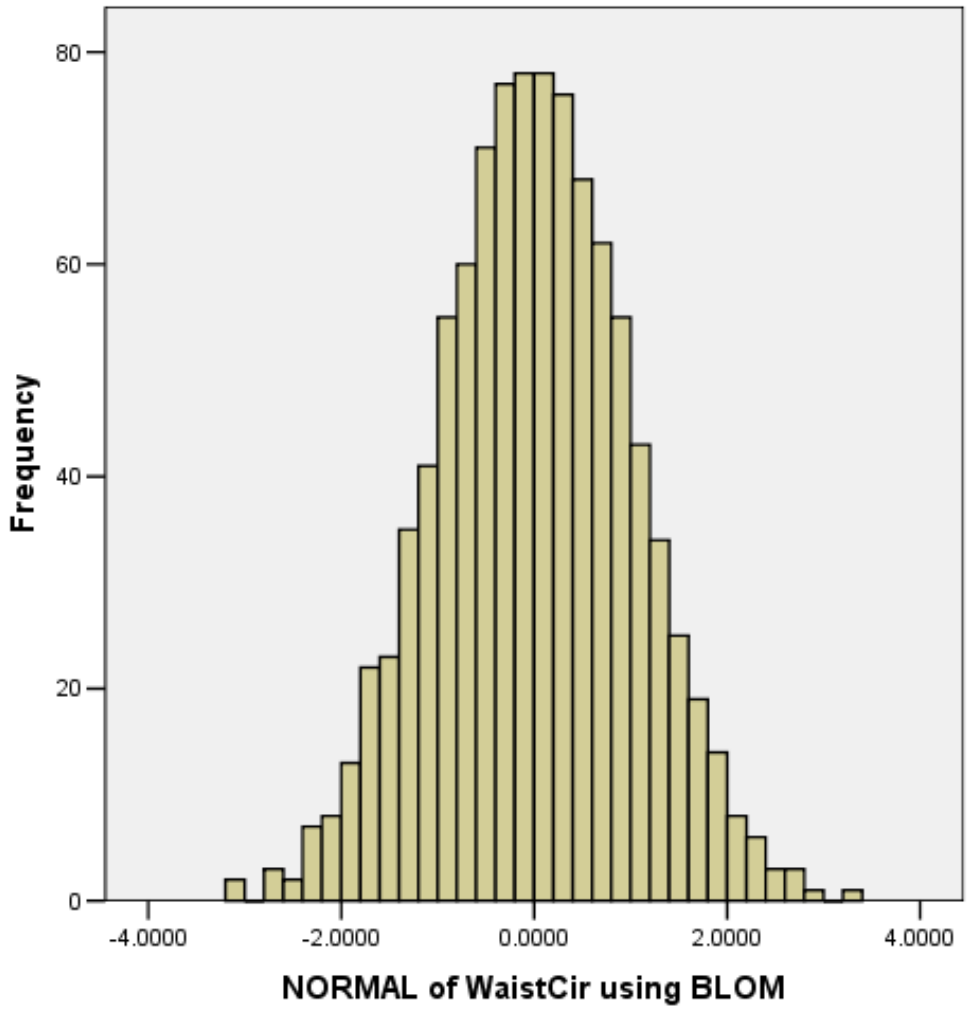

Testiranje hipoteze
TIP PODATKA
Kvalitativni (kategorijski)
1 neovisna varijabla
Goodness of fit x^2
2 ili više neovisnih varijabli
Hi kvadrat (x^2)
2 ili više ovisnih varijabli
McNemar test
Kvantitativni
Povezanost
1 prediktor
Kontinuirana varijabla
Pearson r
Regresija
Rangovi
Spearman r
Više prediktora
Multipla regresija
Razlika
2 skupine
neovisne
t test
Mann-Whitney
ovisne
t test za povezane uzorke
Wilcoxon
Više skupina
neovisne
One-way ANOVA
Kruskal-Wallis
ovisne
ANOVA za ponavljane uzorke
Friedman
Parametrijske
Neparametrijske

Parametrijske metode

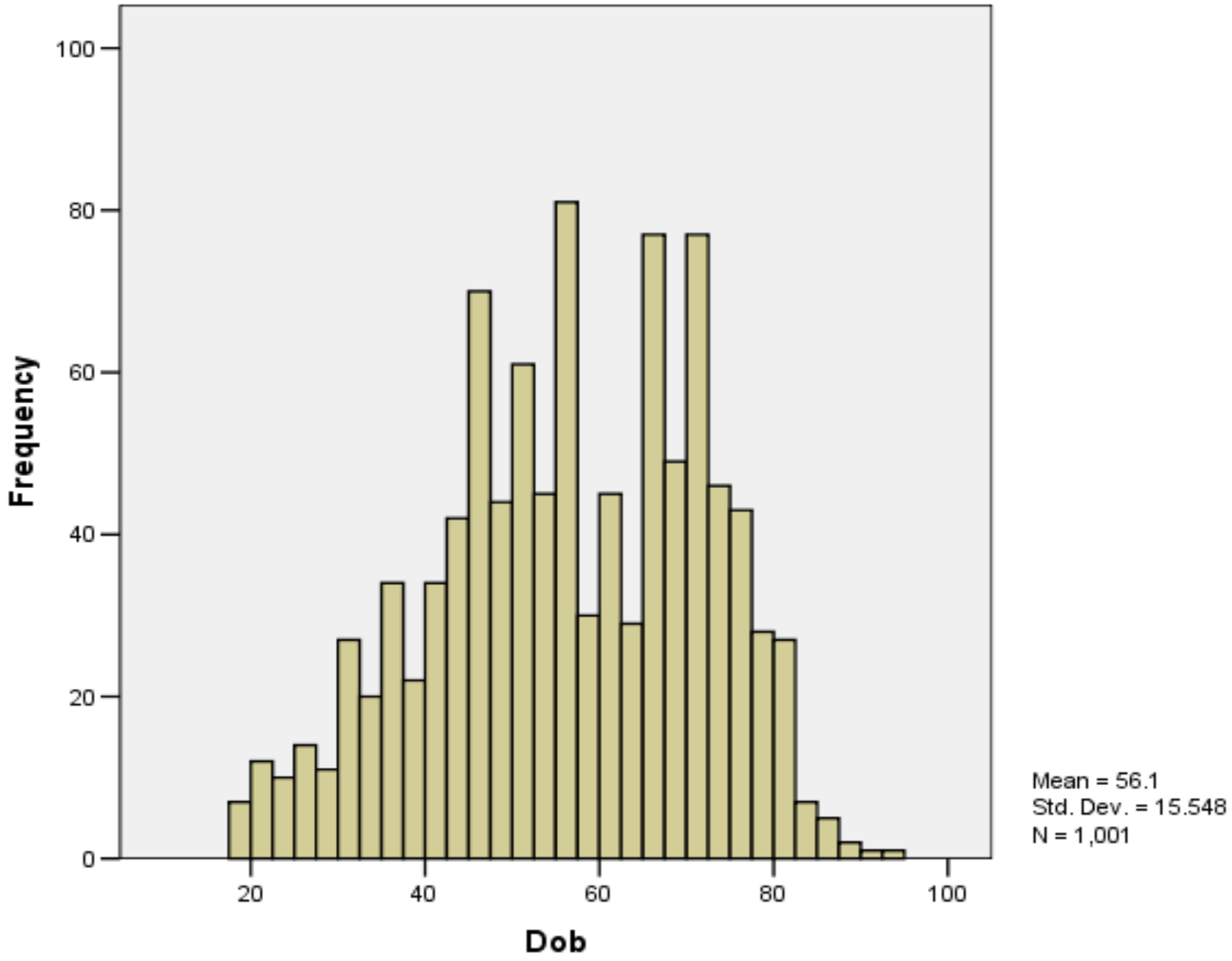
- Temelje se na **parametrima** iz uzorka/populacije
- Zahtijevaju normalnu raspodjelu podataka

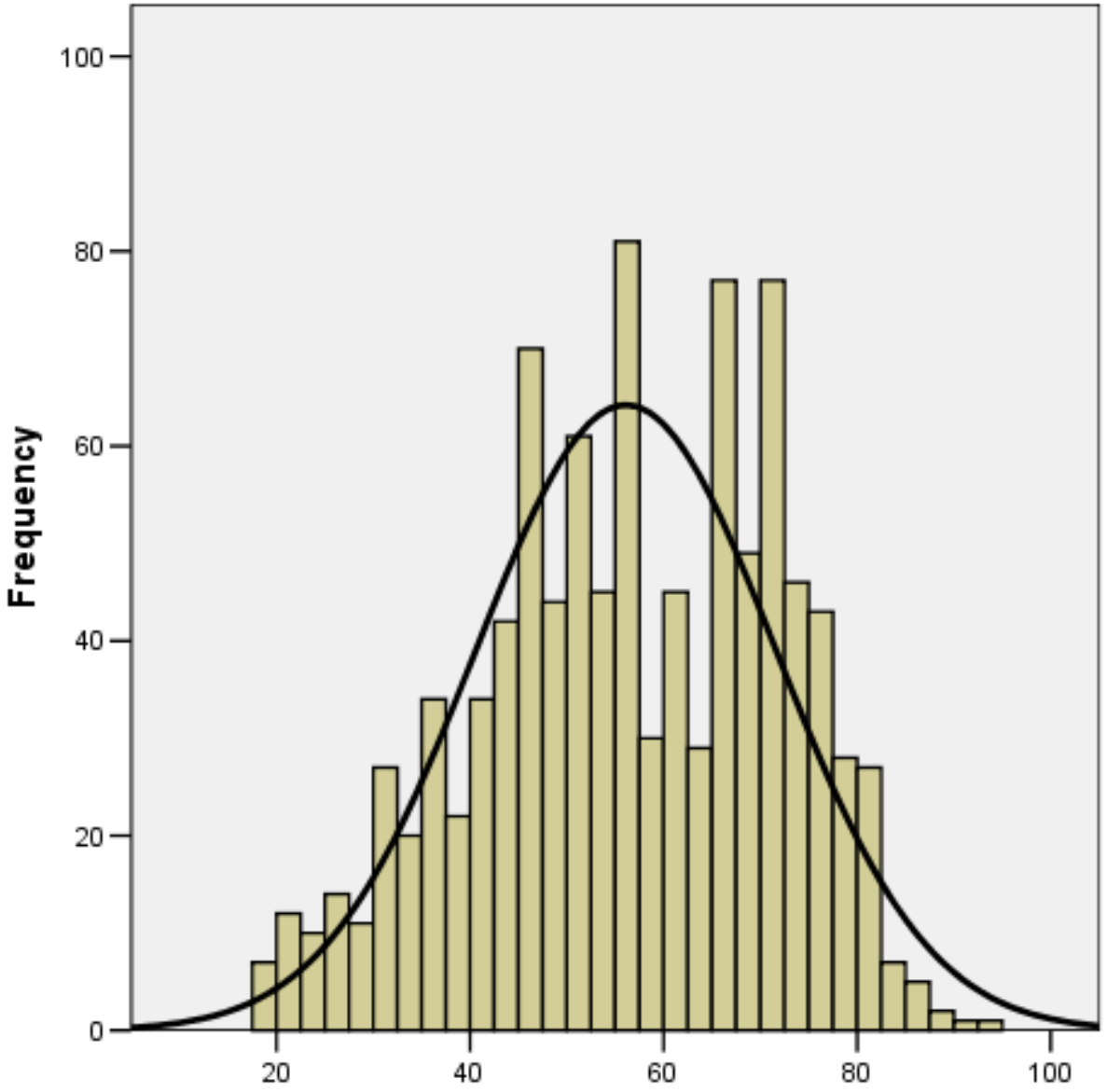


Normalna raspodjela podataka



- "Gaussova"
- eng. *bell shaped*
- Srednja vrijednost ista kao i medijan
- Standardna devijacija određuje širinu


Frequency
100
80
60
40
20
0
20
40
60
80
100
Dob
Mean = 56.1
Std. Dev. = 15.548
N = 1,001



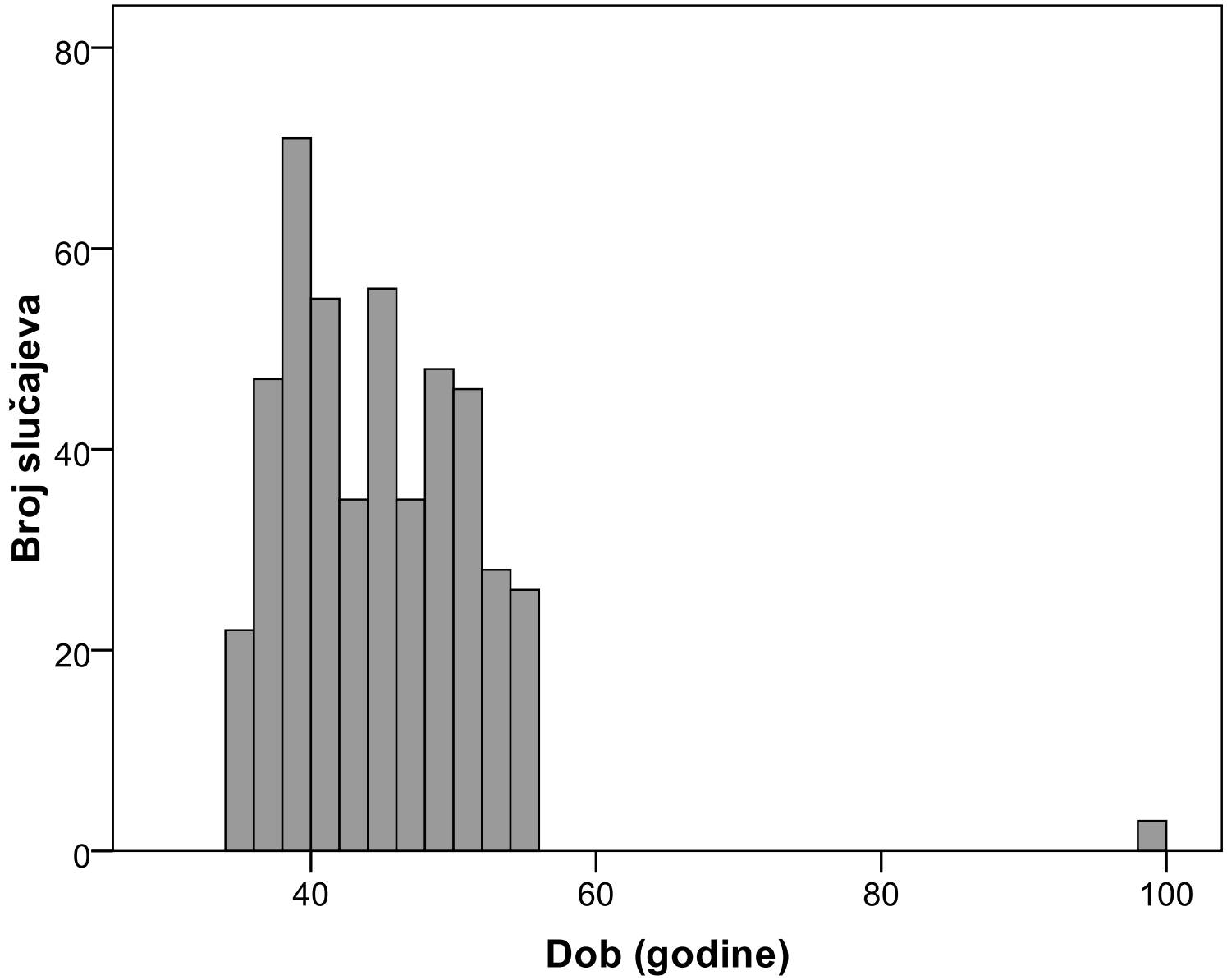
Srednja vrijednost	Standardna devijacija	Medijan	Raspon	Interkvartilni raspon	Najmanja vrijednost	Najveća vrijednost
56.1	15.5	56.0	75.0	24.0	18.0	93.0

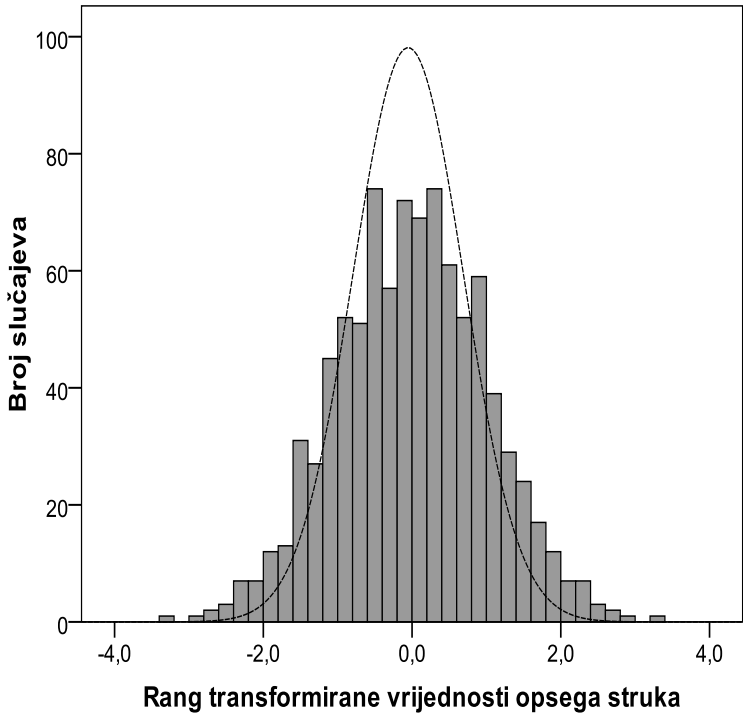
Testiranje normalnosti

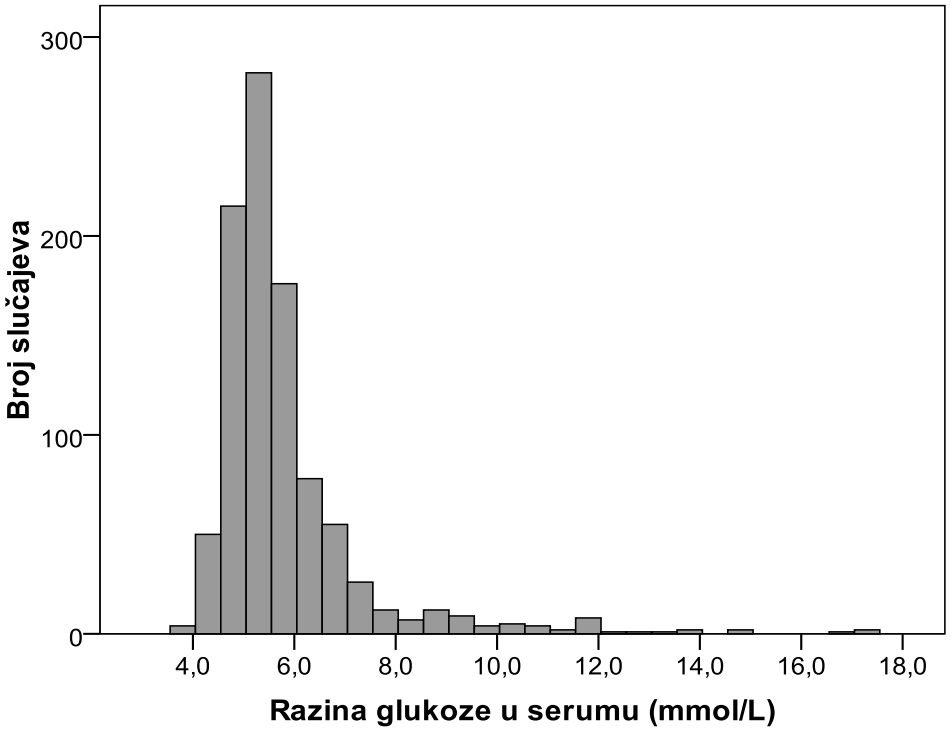
- “Okometrijski”
- Korištenjem posebnih grafičkih prikaza
- Korištenjem statističkih testova
  - Kolmogorov-Smirnov test (>50)
  - Shapiro-Wilk test (<50)

Zašto uopće gledati raspodjelu?

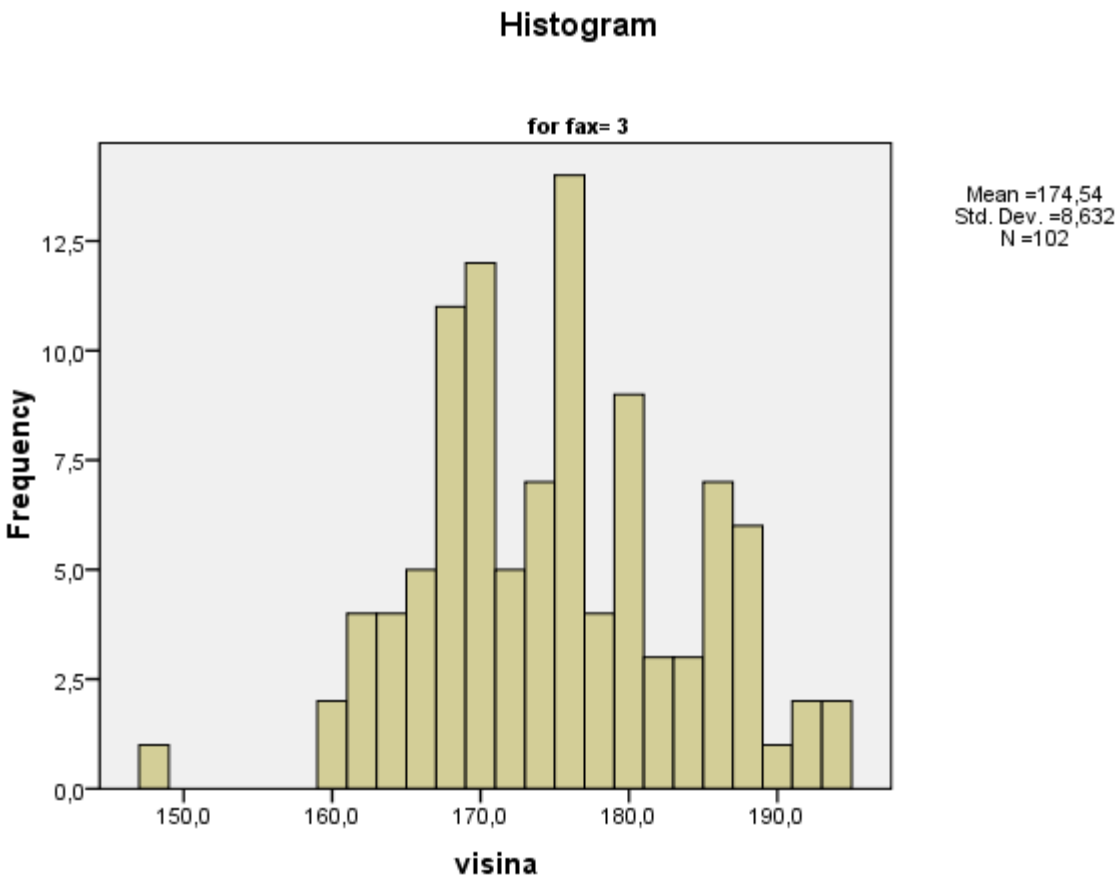
- Zato što o raspodjeli podataka ovisi metoda i tijek analize
- Normalna raspodjela omogućuje upotrebu parametrijskih metoda analize
- Odstupanje od normalne raspodjele onemogućuje upotrebu parametrijskih metoda
- Analiza raspodjele omogućuje uočavanje mogućih pogrešaka u podacima


Broj slučajeva
80
60
40
20
0
40
60
80
100
Dob (godine)


Broj slučajeva
100
80
60
40
20
0
-4,0
-2,0
0,0
2,0
4,0
Rang transformirane vrijednosti opsega struka


Broj slučajeva
300
200
100
0
4,0
6,0
8,0
10,0
12,0
14,0
16,0
18,0
Razina glukoze u serumu (mmol/L)

Normalna raspodjela podataka?



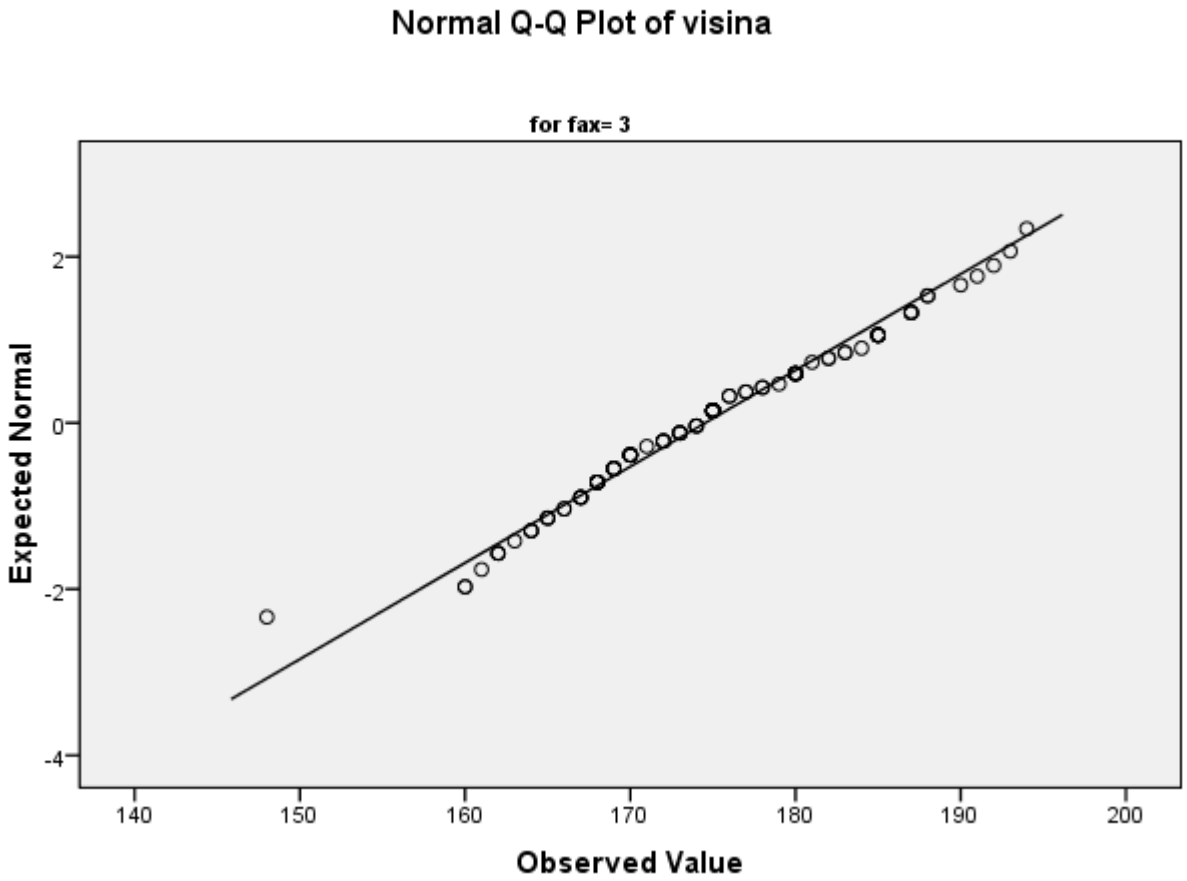
visina Stem-and-Leaf Plot for
fax= 3

Frequency	Stem &	Leaf
1,00	Extremes	(=<148)
3,00	16 .	001
4,00	16 .	2223
6,00	16 .	444555
7,00	16 .	6677777
11,00	16 .	88888899999
8,00	17 .	00000001
8,00	17 .	22223333
15,00	17 .	444555555555555
4,00	17 .	6677
3,00	17 .	889
9,00	18 .	000000001
4,00	18 .	2233
8,00	18 .	45555555
4,00	18 .	7777
2,00	18 .	88
2,00	19 .	01
2,00	19 .	23
1,00	19 .	4

Stem width: 10,0
Each leaf: 1 case(s)

Normalna raspodjela podataka?

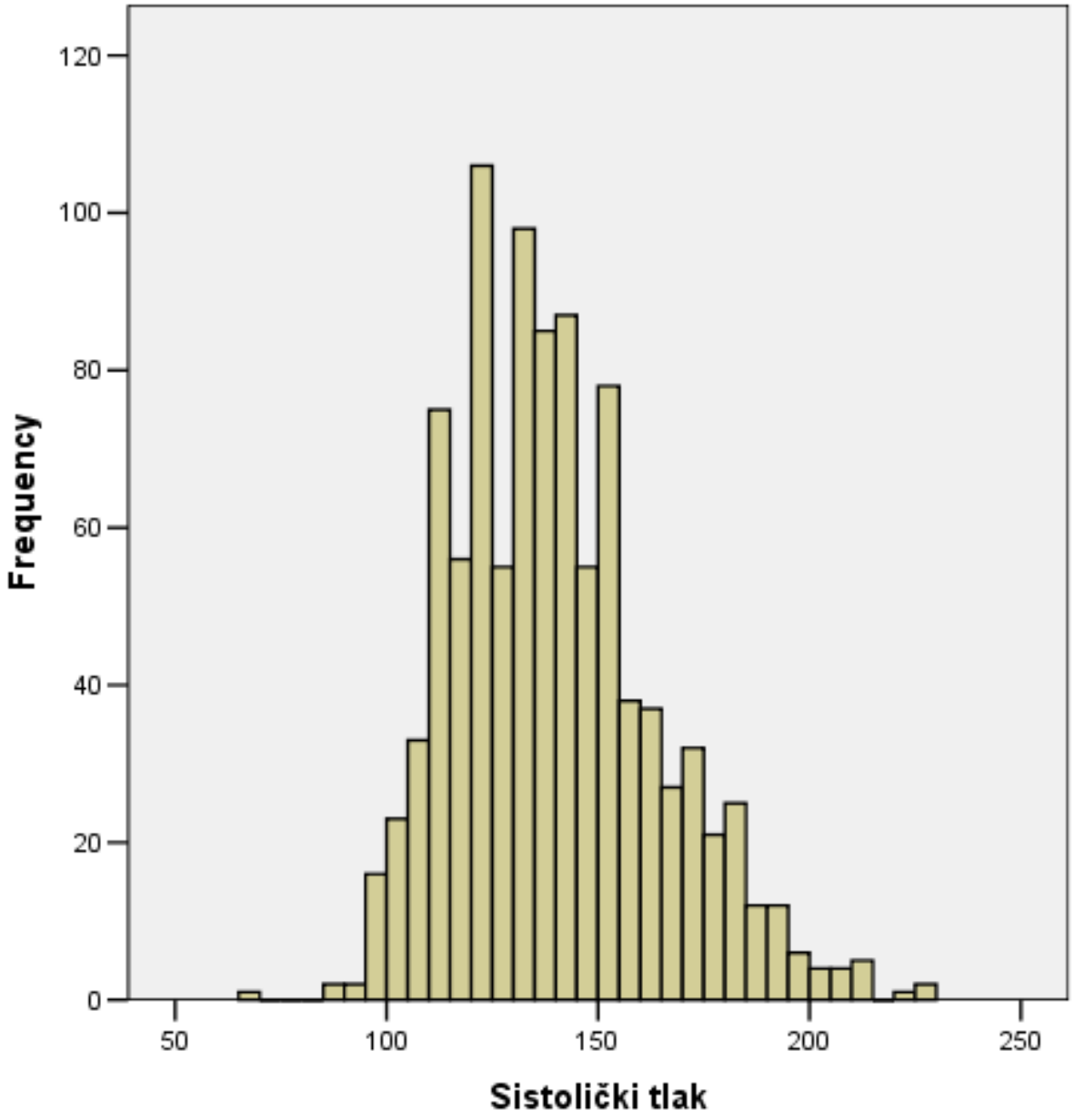
Normalna raspodjela podataka?



Normalna raspodjela podataka?

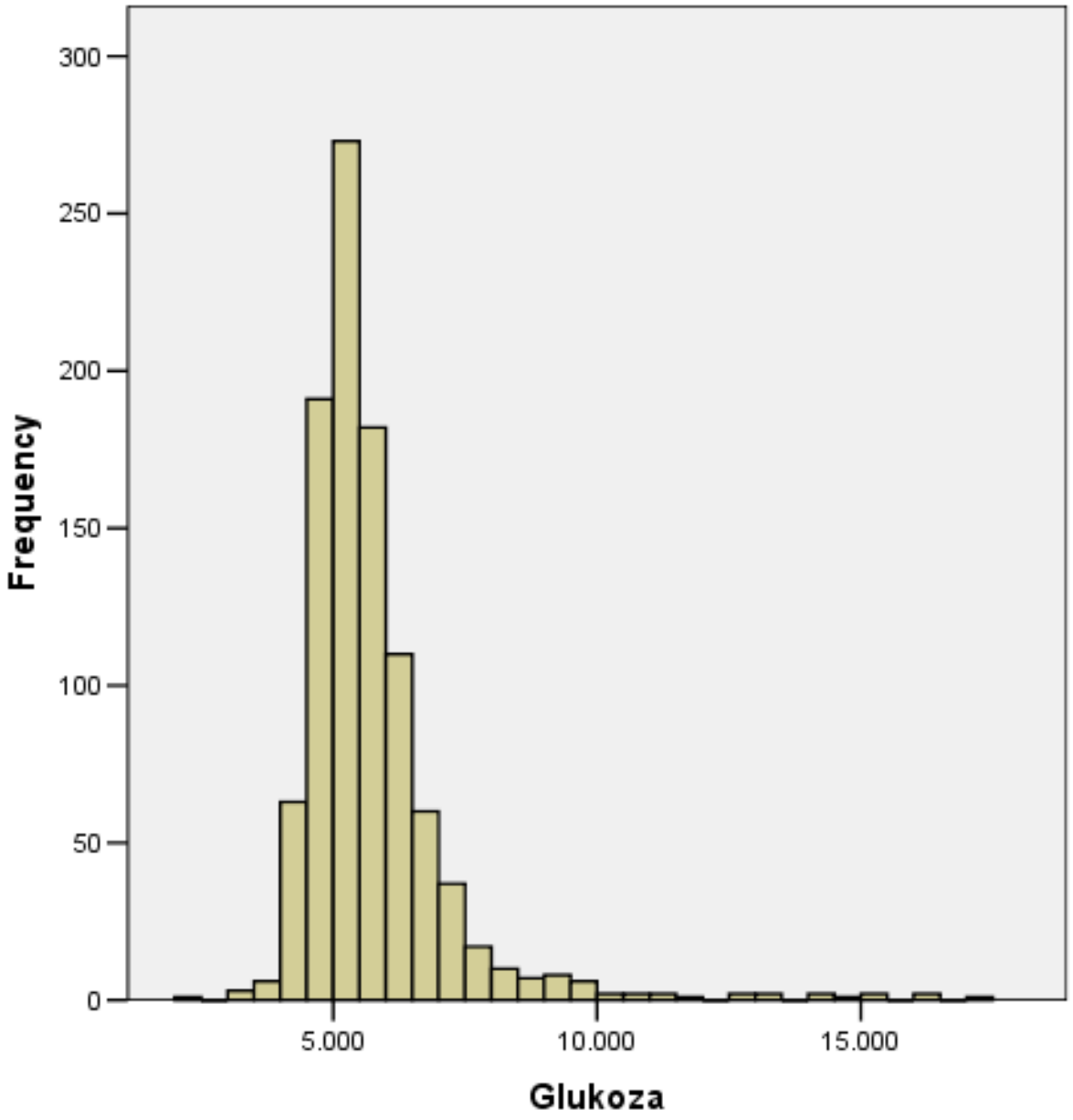
Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	fax	Statistic	df	Sig.	Statistic	df	Sig.
visina	3	,096	102	,021	,982	102	,180



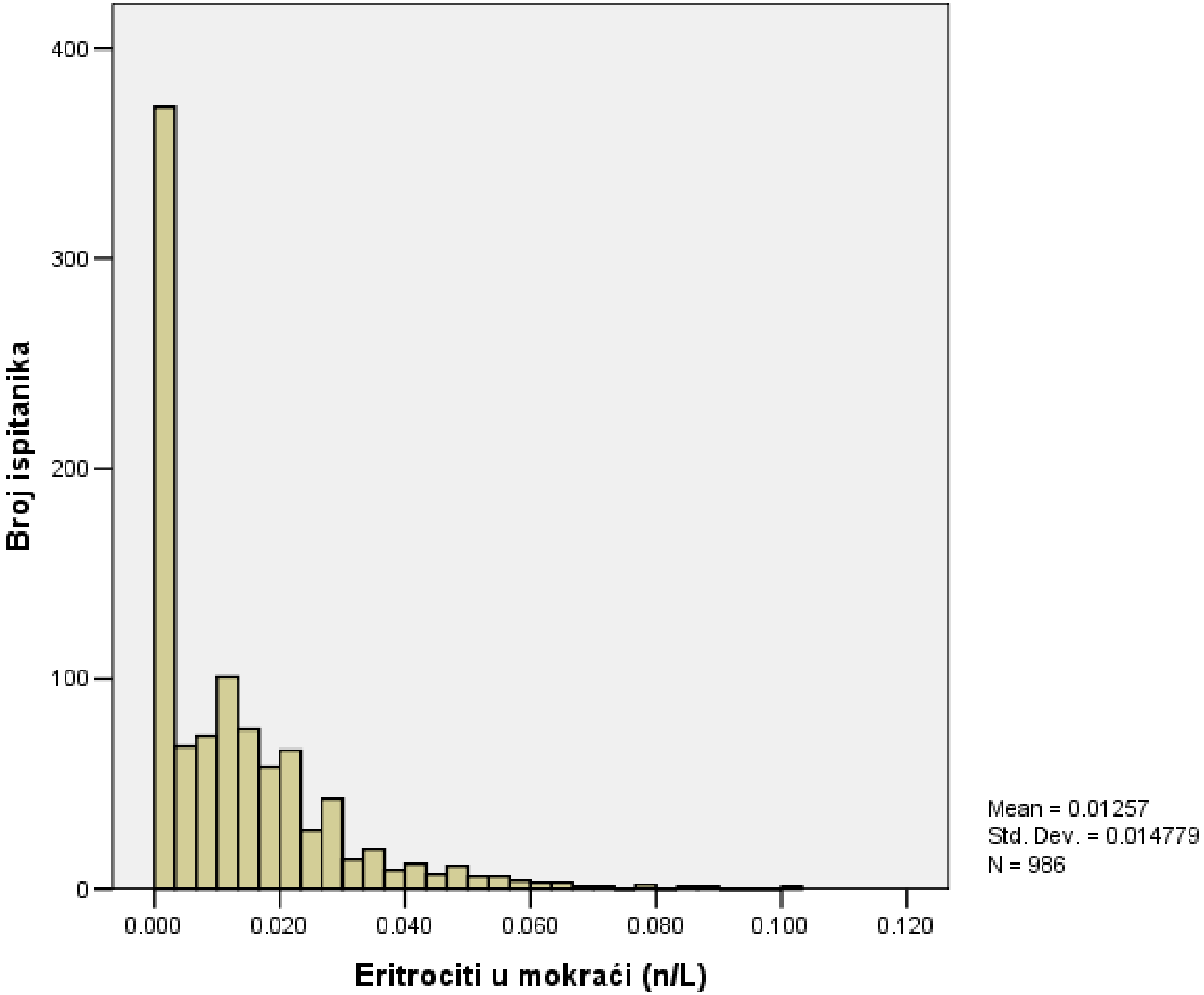
Aritmetička sredina	**138.3**
Std. Devijacija	**24.1**
Medijan	**135.0**
Min	**69.0**
Max	**230.0**
Raspon	**161.0**
Interkvartilni raspon	**32.0**

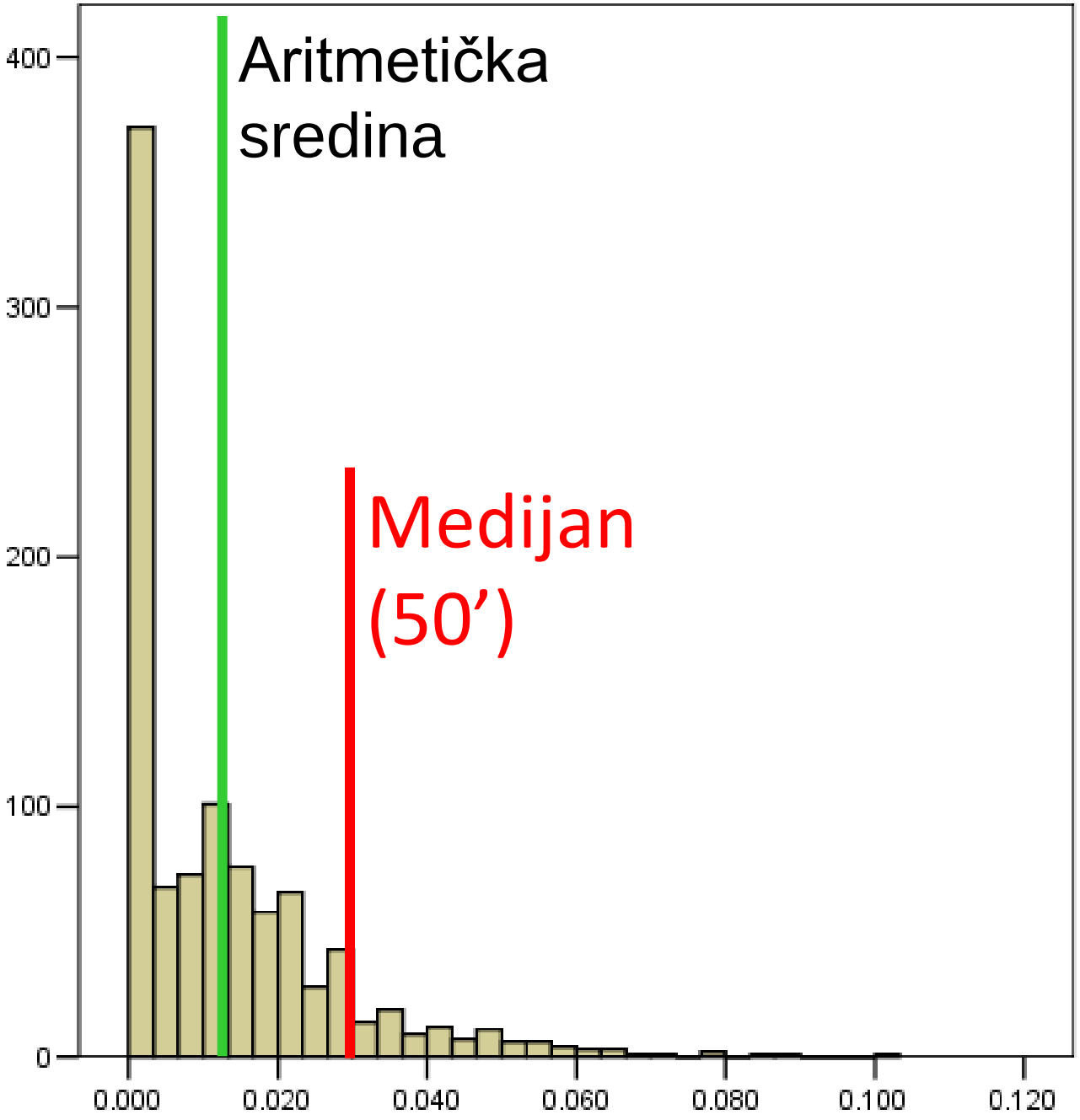
Mean = 138.28
Std. Dev. = 24.136
N = 998



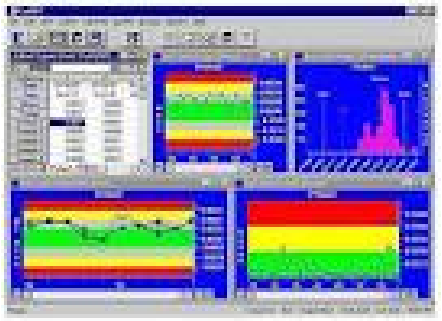
Aritmetička sredina	**5.69**
Std. Devijacija	**1.48**
Medjian	**5.40**
Min	**2.30**
Max	**17.40**
Raspon	**15.10**
Interkvartilni raspon	**1.10**

Mean = 5.69416
Std. Dev. = 1.476751
N = 993


400
300
200
100
0
Broj ispitanika
0.000
0.020
0.040
0.060
0.080
0.100
0.120
Eritrociti u mokraći (n/L)
Mean = 0.01257
Std. Dev. = 0.014779
N = 986

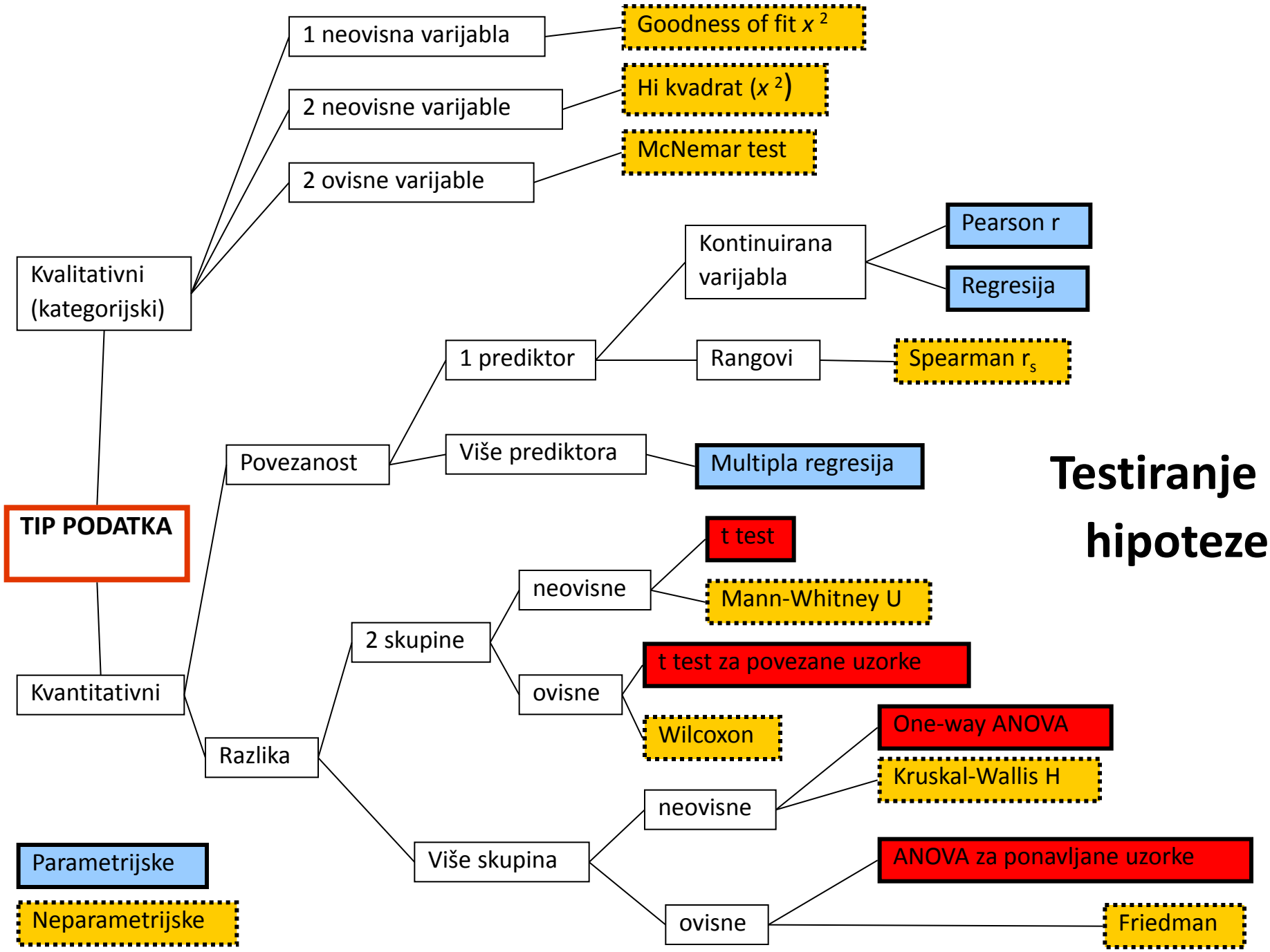

Aritmetička sredina
Medijan (50’)
400
300
200
100
0
0.000
0.020
0.040
0.060
0.080
0.100
0.120

Provedba statističke raščlambe podataka



Odabir statističkog testa

- 2 nepovezane skupine: t-test (engl. *independent samples t-test*)
- 2 povezane skupine: t-test za povezane uzorke (engl. *dependent samples t-test*)
- Više od 2 neovisna uzorka: F-test ili ANOVA (analysis of variance)+*post-hoc* test
- Više od 2 ovisna uzorka: faktorska ANOVA i AUC


Testiranje hipoteze
TIP PODATKA
Kvalitativni (kategorijski)
1 neovisna varijabla
Goodness of fit x^2
2 neovisne varijable
Hi kvadrat (x^2)
2 ovisne varijable
McNemar test
Kvantitativni
Povezanost
1 prediktor
Kontinuirana varijabla
Pearson r
Regresija
Rangovi
Spearman r_s
Više prediktora
Multipla regresija
Razlika
2 skupine
neovisne
t test
Mann-Whitney U
ovisne
t test za povezane uzorke
Wilcoxon
Više skupina
neovisne
One-way ANOVA
Kruskal-Wallis H
ovisne
ANOVA za ponavljane uzorke
Friedman
Parametrijske
Neparametrijske

Cilj analize	Vrsta varijable		
	Numerička, postoji normalna raspodjela (parametrijske metode)	Ordinalna ili numerička, čija raspodjela odstupa od normalne (neparametrijske metode)	Kategorijska
Usporedba dvije skupine neovisnih podataka	t-test za neovisne uzorke	Mann-Whitneyev test	Hi-kvadrat test (Fisherov egzaktni test)
Usporedba dvije skupine povezanih podataka	t-test za povezane uzorke	Wilcoxonov test	McNemarov test
Usporedba tri ili više skupina neovisnih podataka	Analiza varijance (ANOVA)	Kruskal-Wallisov test	Hi-kvadrat test
Usporedba tri ili više skupina povezanih podataka	Ponavljana ANOVA	Friedmanov test	Cochraneov Q test
Korelacija	Pearsonova korelacija	Spearmanova korelacija	Koeficijent kontingencije
Predviđanje jedne ovisne varijable (engl. *outcome or dependent variable*) na temelju jedne ili više prediktorskih varijabli	Linearna regresija	Neparametrijska ili ordinalna regresija	Logistička regresija

t-test za neovisne uzorke

- 1908 William Sealy Gosset
- Osmislio je novi test za kontrolu kvalitete piva u pivovari Guinness
- Objavio rezultate u časopisu Biometrika, ali nije mogao koristiti svoje ime zbog očuvanja poslovne tajne

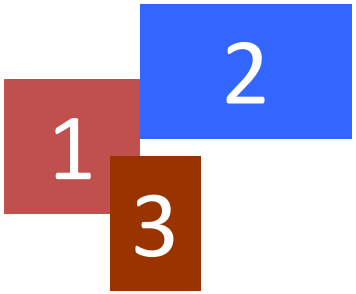
© Guinness

t-test za ovisne uzorke

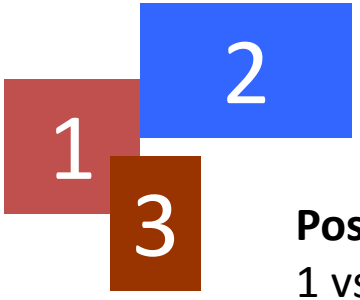
- Podaci koji su povezani
- Npr. dužina lijeve i desne očne jabučice
- Interpretacija ista kao i t-test za neovisne uzorke

ANOVA

- *Analysis of variance*
- Više nepovezanih skupina
- Međutim, sam test nije dovoljan (P<0,001)
- *Post-hoc* test – usporedba svake skupine sa svakom



ANOVA P<0,001



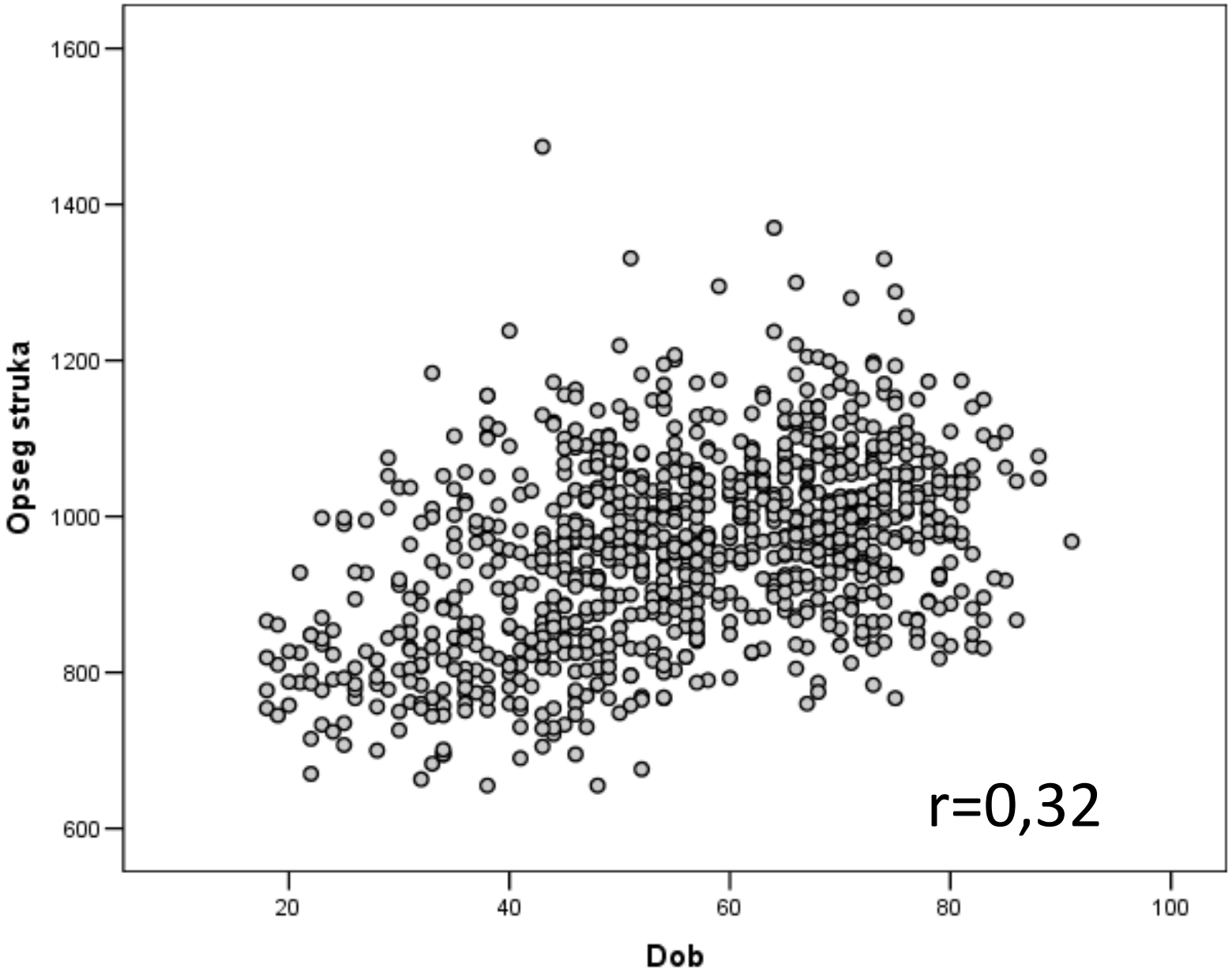
Post-hoc
1 vs. 2 P=0,621
1 vs. 3 P=0,003
2 vs. 3 P<0,001

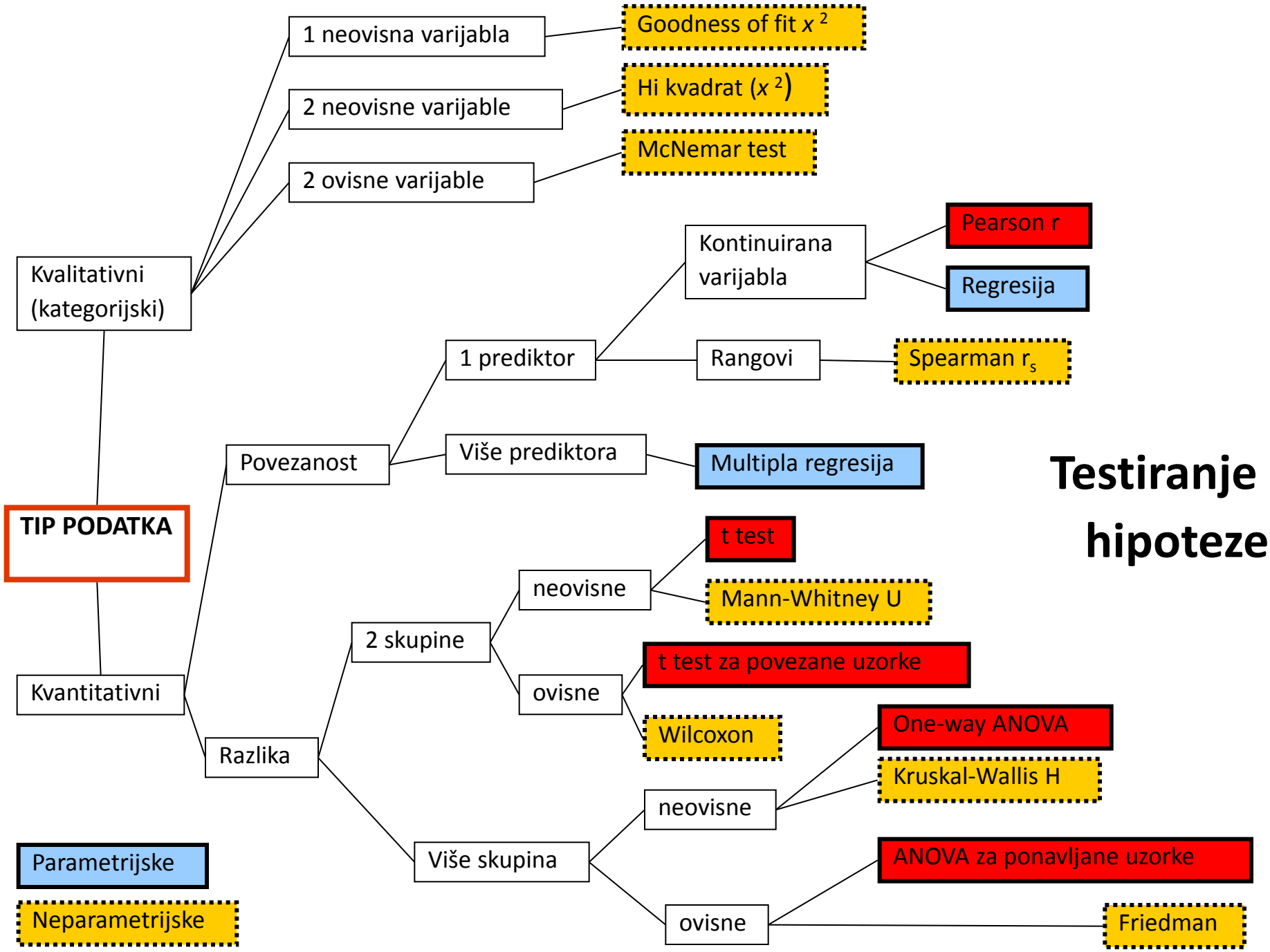
Parametrijska korelacija

- Povezanost dvije kontinuirane numeričke varijable koje obilježava normalnu raspodjelu podataka
- Koeficijent korelacije (r, *test statistic*)
- Govori o snazi povezanosti, a kreće se od -1,0 do 1,0
- Koeficijent korelacije r=0 ukazuje na nepostojanje korelacije

Pearsonov test korelacije

- Parametrijski model korelacije
- Zasniva se na testiranju snage povezanosti dvije varijable
- Uvijek prikazujte graf rasapa (*scatterplot*) na kojem se vidi priroda korelacije
- Korelacija mora biti linearna


Opseg struka
1600
1400
1200
1000
800
600
20
40
60
80
100
Dob
r=0,32


Testiranje hipoteze
TIP PODATKA
Kvalitativni (kategorijski)
1 neovisna varijabla
Goodness of fit x^2
2 neovisne varijable
Hi kvadrat (x^2)
2 ovisne varijable
McNemar test
Kvantitativni
Povezanost
1 prediktor
Kontinuirana varijabla
Pearson r
Regresija
Rangovi
Spearman r_s
Više prediktora
Multipla regresija
Razlika
2 skupine
neovisne
t test
Mann-Whitney U
ovisne
t test za povezane uzorke
Wilcoxon
Više skupina
neovisne
One-way ANOVA
Kruskal-Wallis H
ovisne
ANOVA za ponavljane uzorke
Friedman
Parametrijske
Neparametrijske

GIGO

- *Garbage in, garbage out*
- Niti najbolja statistička obrada neće popraviti loš dizajn istraživanja, unos podataka ili loše istraživačko pitanje

- Asking a statistician to help after the experiment has been completed is like talking to a pathologist. It is then that the statistician can tell you what the project died of.

Sir Ronald Aylmer Fisher (1890-1962)

http://en.wikipedia.org/wiki/Ronald_Fisher

Regresija: krvni tlak i prihodi

- Postoji jasna i očita povezanost krvnog tlaka i razine prihoda, na način da u uzorku ispitanici sa najvišim primanjima imaju najviši krvni tlak
- Objašnjenje?

Tko ima visoke prihode?

- Stariji
- Muškarci
- Višeg stupnja obrazovanja*

Regresija

- Povezanost više (prediktorskih) varijabli sa jednom ciljnom (ovisnom) varijablom
- Npr. utjecaj spola i dobi na visinu krvnog tlaka
- Istovremeni prikaz utjecaja više varijabli na jednu
- Epidemiološki rečeno, ovo su varijable posredne povezanosti (engl. *confounding*)
- Kako se riješiti ovog učinka?

Oblici regresijske analize

- Linearna i multipla linearna
- Logistička
- Cox (*hazard model*)
- Ordinalna

Pretpostavke korištenja linearne regresije

- Ciljna varijabla ima normalu raspodjelu
- Prediktorske varijable imaju normalnu raspodjelu
- Prediktorske varijable ne mogu biti ordinalne ili kategorijske (samo binarne i kontinuirane normalne)

Logistička regresija

- Ciljna varijabla je binarna (npr. zdrav-bolestan, živ-mrtav, …)
- Mjeri utjecaj pojedine klase prediktorske varijable na ishod

Prosjek ocjena

- Ocjene od 1-5 su brojčane
- Međutim, one su **diskretne**, nisu kontinuirane
- Ovaj tip podatka nikako ne može imati normalnu raspodjelu

Neparametrijske metode

- Analitičke metode koje se ne zasnivaju na pretpostavci raspodjele podataka
- **NEMA** srednje vrijednosti i standardne devijacije
- Podaci su po svojoj prirodi nominalni ili ordinalni

Prednost NP metoda

- Mogućnost analize raznolikih uzoraka (engl. *outliers*)
- Analiza se svodi na rang podataka ne na stvarne vrijednosti

Kada obavezno NP metode?

- Mali uzorci (N<30)
- Varijable koje nemaju normalnu raspodjelu (npr. enzimi, biokemijski pokazatelji, krvni tlak, …)
- Ordinalne varijable (ocjene, starost u godinama, …)

Zašto ne koristiti NP metode?

- Otežana interpretacija (medijan i raspon)
- Ponekad nemoguće pokazati razliku dvije varijable (iste vrijednosti medijana)
- Smanjena statistička snaga testa i povećana šansa za pogreške

Pogreške u analizi

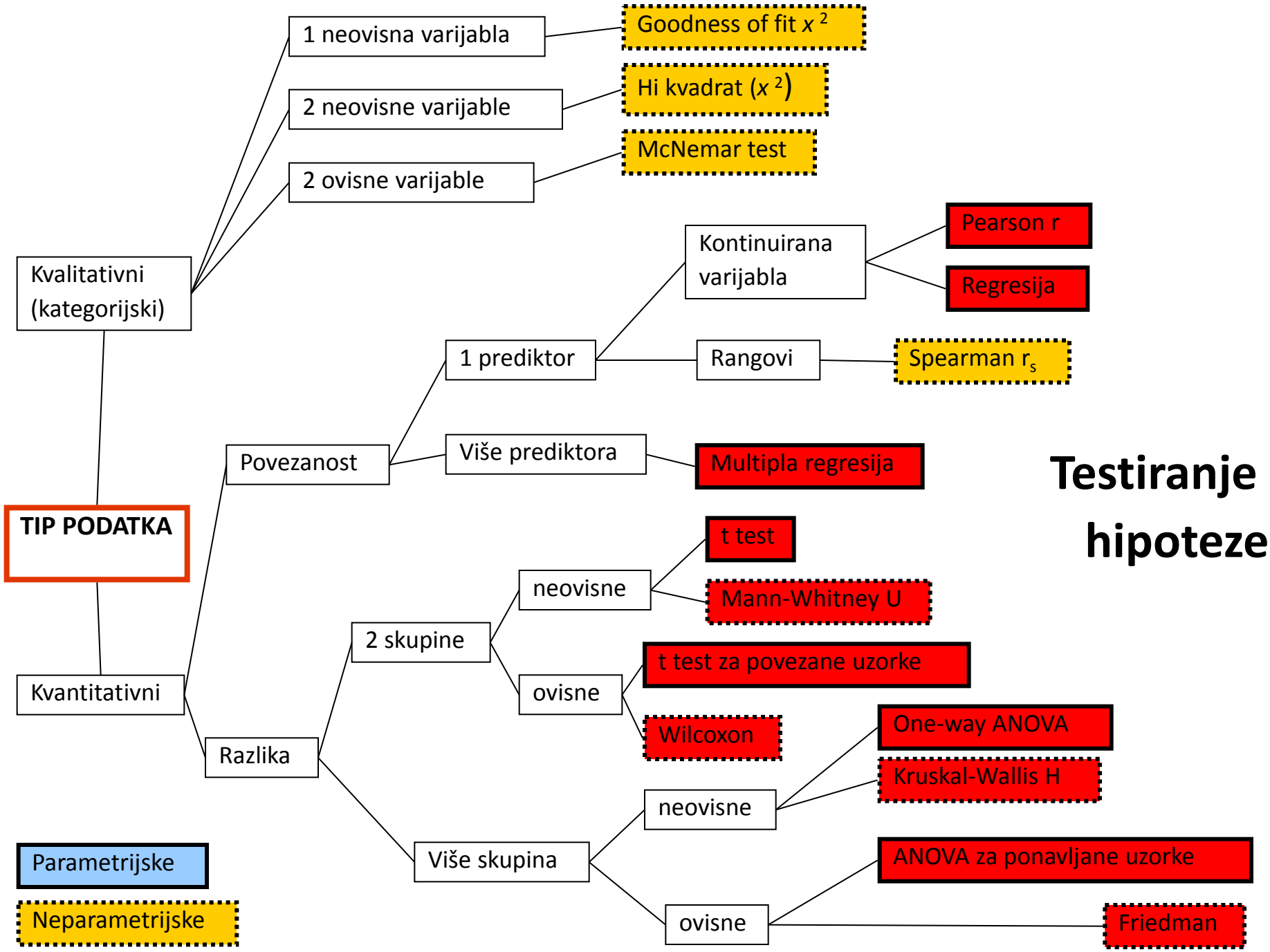
- Pogreška tipa I: lažno pozitivni rezultat
  - Odbijanje nul-hipoteze kada je ona stvarno istina, tj. prikazivanje rezultata kao statistički značajan kada on uistinu nije
- Pogreška tipa II: promašaj stvarnog učinka
  - Pogreška koja nastaje jer se ne odbacuje nul-hipoteza kada je ona lažna, tj. odbacivanje stvarnog rezultata i proglašavanje neznačajnim

Neparametrijske metode

- 2 neovisna uzorka – Mann-Whitney (*t-test*)
- 2 povezana uzorka – Wilcoxon (*t-test PU*)
- Više od dva neovisna uzorka – Kruskal-Wallis (*ANOVA*)
- Više od dva povezana uzorka – Friedman (*faktorska ANOVA*)

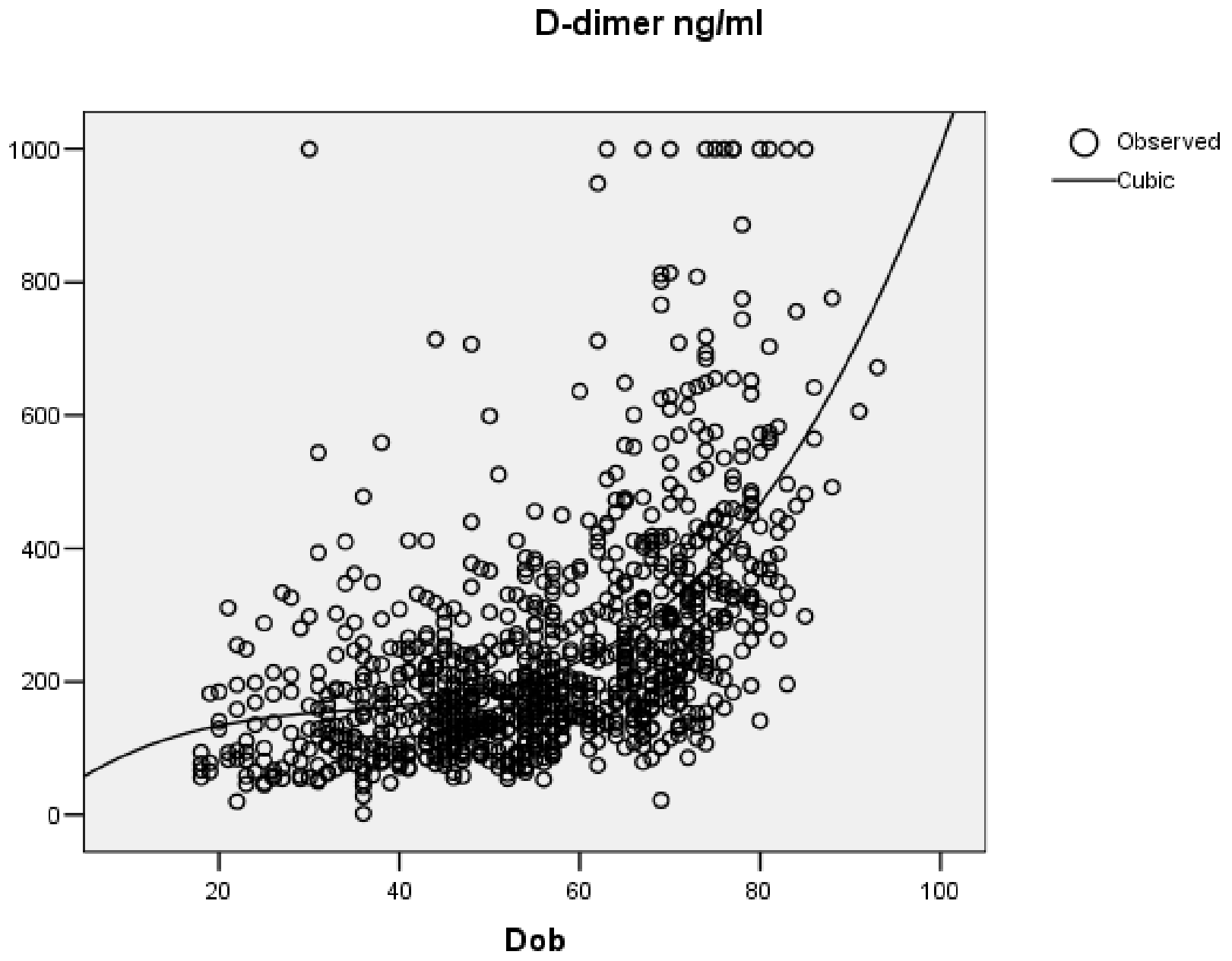
Usporedba P i NP metoda

- Moguće je izračunati rezultata i P i NP metoda za neki uzorak i usporediti značajnosti
- Oba uzorka isto – jednostavno
- Problem – P i NP rezultati se razlikuju


Testiranje hipoteze
TIP PODATKA
Kvalitativni (kategorijski)
1 neovisna varijabla
Goodness of fit x^2
2 neovisne varijable
Hi kvadrat (x^2)
2 ovisne varijable
McNemar test
Kvantitativni
Povezanost
1 prediktor
Kontinuirana varijabla
Pearson r
Regresija
Rangovi
Spearman r_s
Više prediktora
Multipla regresija
Razlika
2 skupine
neovisne
t test
Mann-Whitney U
ovisne
t test za povezane uzorke
Wilcoxon
Više skupina
neovisne
One-way ANOVA
Kruskal-Wallis H
ovisne
ANOVA za ponavljane uzorke
Friedman
Parametrijske
Neparametrijske

Neparametrijska korelacija

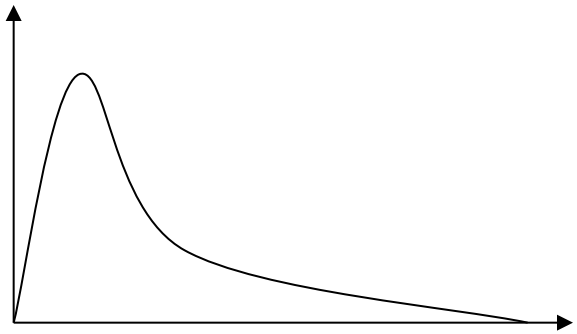
- Korelacija dvije kvantitativne kontinuirane varijable koje nisu povezane linearno ili nemaju normalnu raspodjelu podataka
- Ordinalne varijable
- Spearmanov rank test
- Isti pokazatelji kao i Pearsonov test (r, P vrijednost)


D-dimer ng/ml
1000
800
600
400
200
0
20
40
60
80
100
Dob
Observed
Cubic

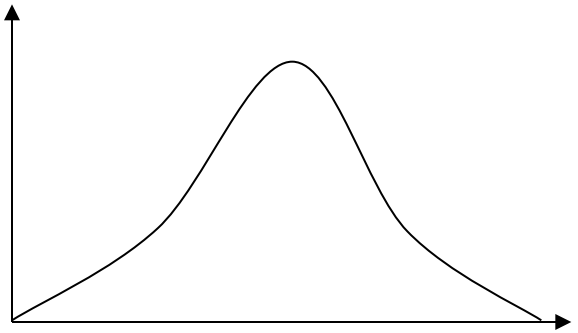
Što sa raspodjelom podataka?

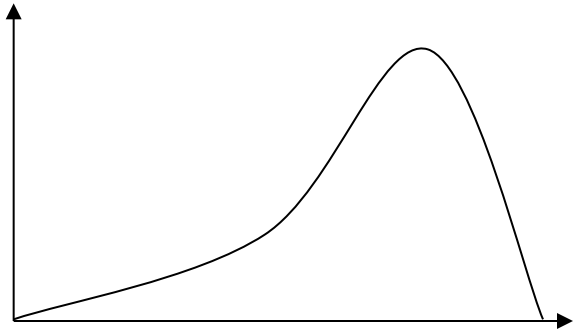
- Parametrijske metode?
- Neparametrijske metode?
- Regresija?
- Transformacija podataka – računska operacija s podacima koja rezultira promjenom raspodjele podataka

Oblici transformacije podataka



Logaritamska transformacija [log(x)]





Kvadratična transformacija (x^2)

Table 3. Effect of relative heterozygosity (h_R) on QT values[a]

QT	N	Multiple regression model		
		Coefficient	Standard error	P-value
SBP	223	−102.8	42.0	**0.015**
DBP	223	−47.7	20.5	**0.021**
Log(total cholesterol)	200	−1.083	0.439	**0.014**
Log(LDL cholesterol)	201	−1.539	0.597	**0.011**
Log(HDL cholesterol)	200	+0.172	0.368	0.640
Log(triglyceride)	201	−0.461	0.914	0.615
Log{log(glucose)}	201	+0.241	0.216	0.266
Log(creatinine)[b]	200	−0.084	0.328	0.799
Urate	201	−147.9	136.7	0.281
Forced vital capacity	201	−2.381	1.250	0.058
Forced expiratory flow (FEF_{25})[b]	200	−3.174	1.366	**0.021**
Forced expiratory flow (FEF_{50})	201	−1.433	2.582	0.580
Forced expiratory flow (FEV_1)	201	−1.458	1.065	0.173
Peak expiratory flow	201	+4.257	3.489	0.224
Log(body mass index)[c]	231	+0.068	0.279	0.807

Carothers AD, Rudan I, Kolcic I, Polasek O, Hayward C, Wright AF, Campbell H, Teague P, Hastie ND, Weber JL. Estimating human inbreeding coefficients: comparison of genealogical and marker heterozygosity approaches. Annals of Human Genetics 2006;70(5):666-76.

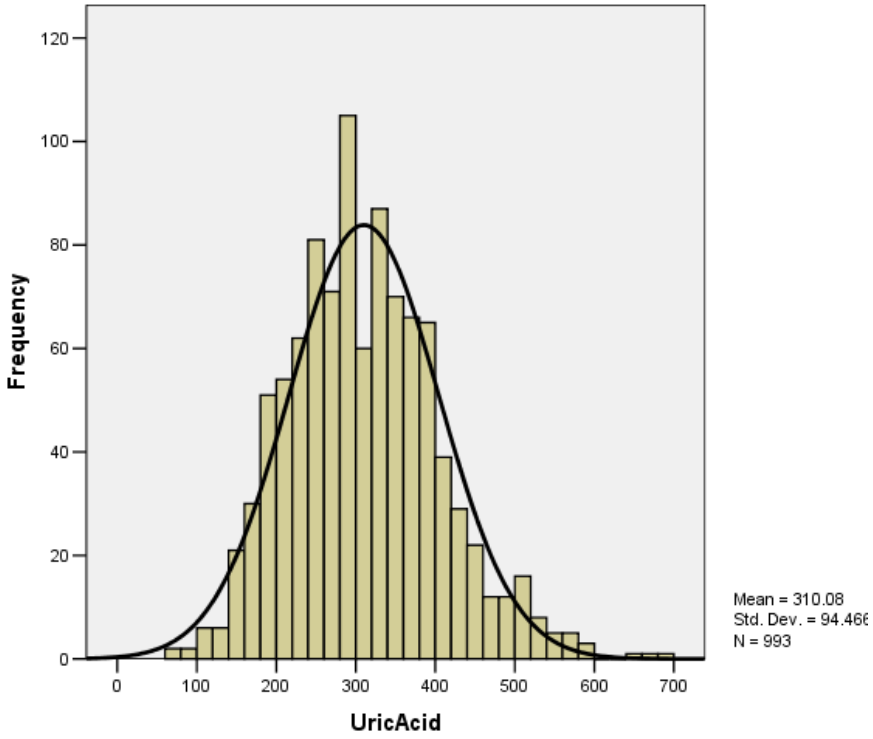
Oblici transformacija

- Logaritamska
- Kvadratična
- Korjenska
- Inverzna
- Logit (proporcije)

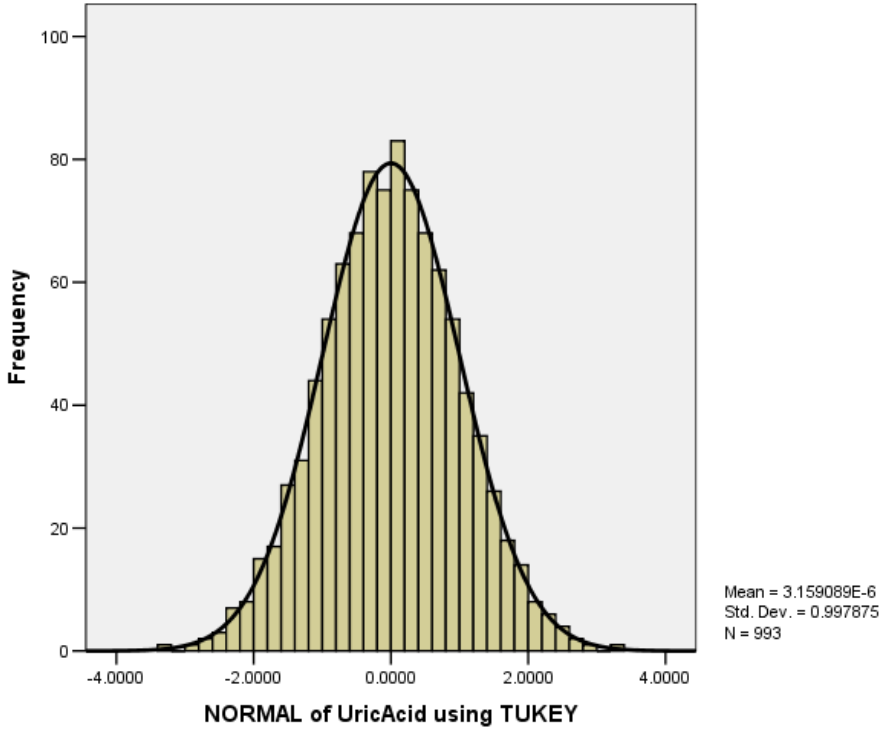
Rang-normalnost transformacija

- Transformacija koja rangira sve uzorke (slaže po redu), a zatim njihove rangove zamjenjuje za vrijednosti dobivene iz izračuna normalne raspodjele podataka na temelju parametara uzorka

Rang-normalnost transformacija

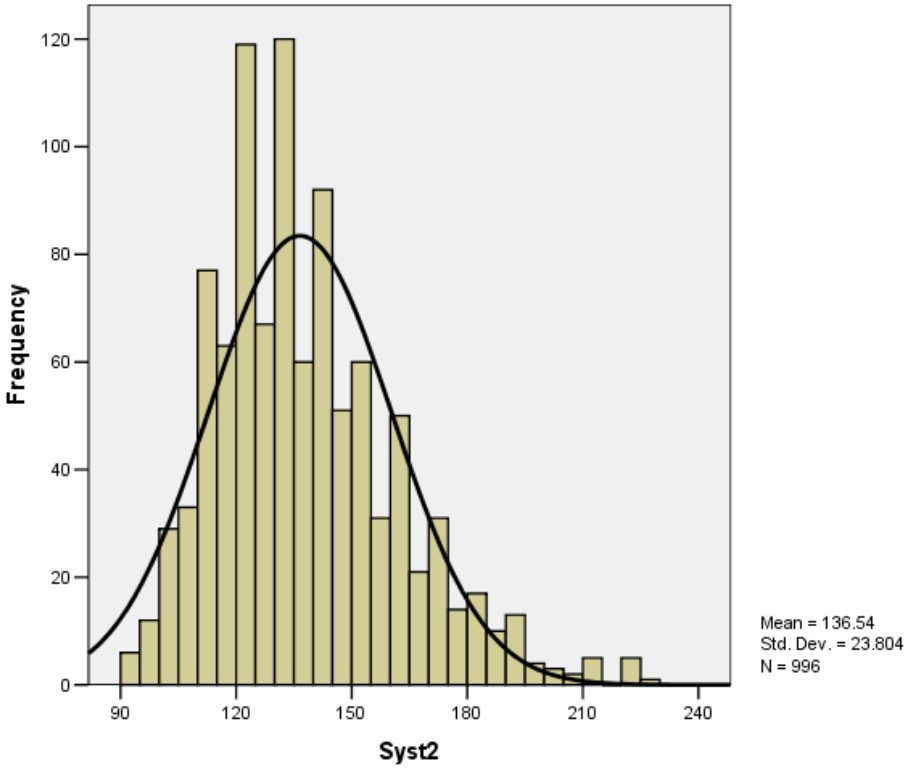


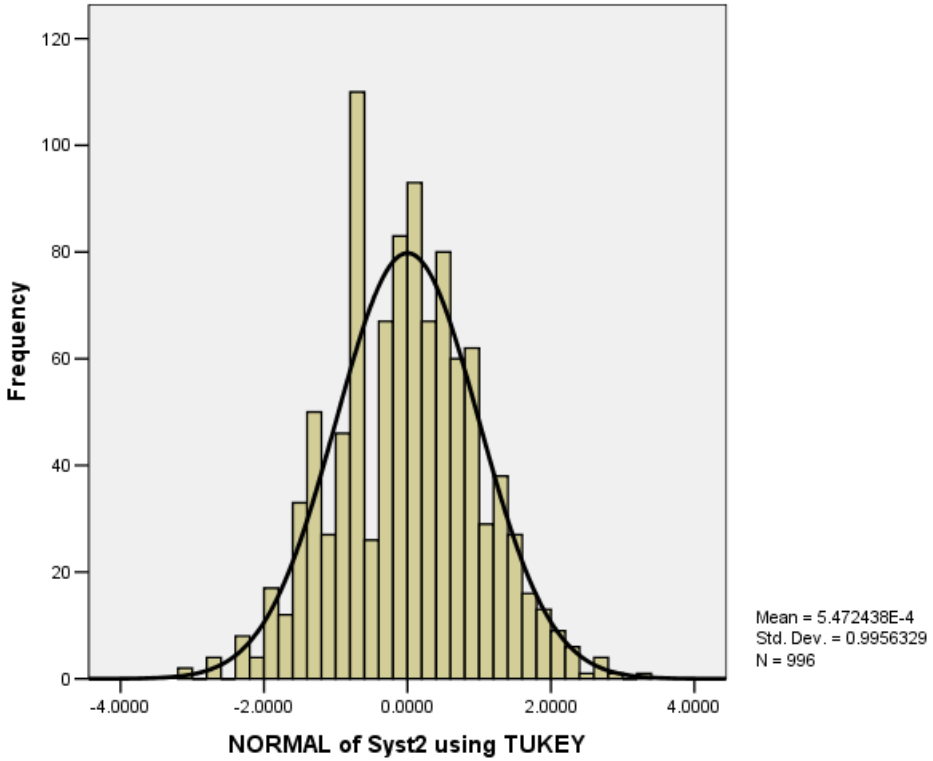
Prije

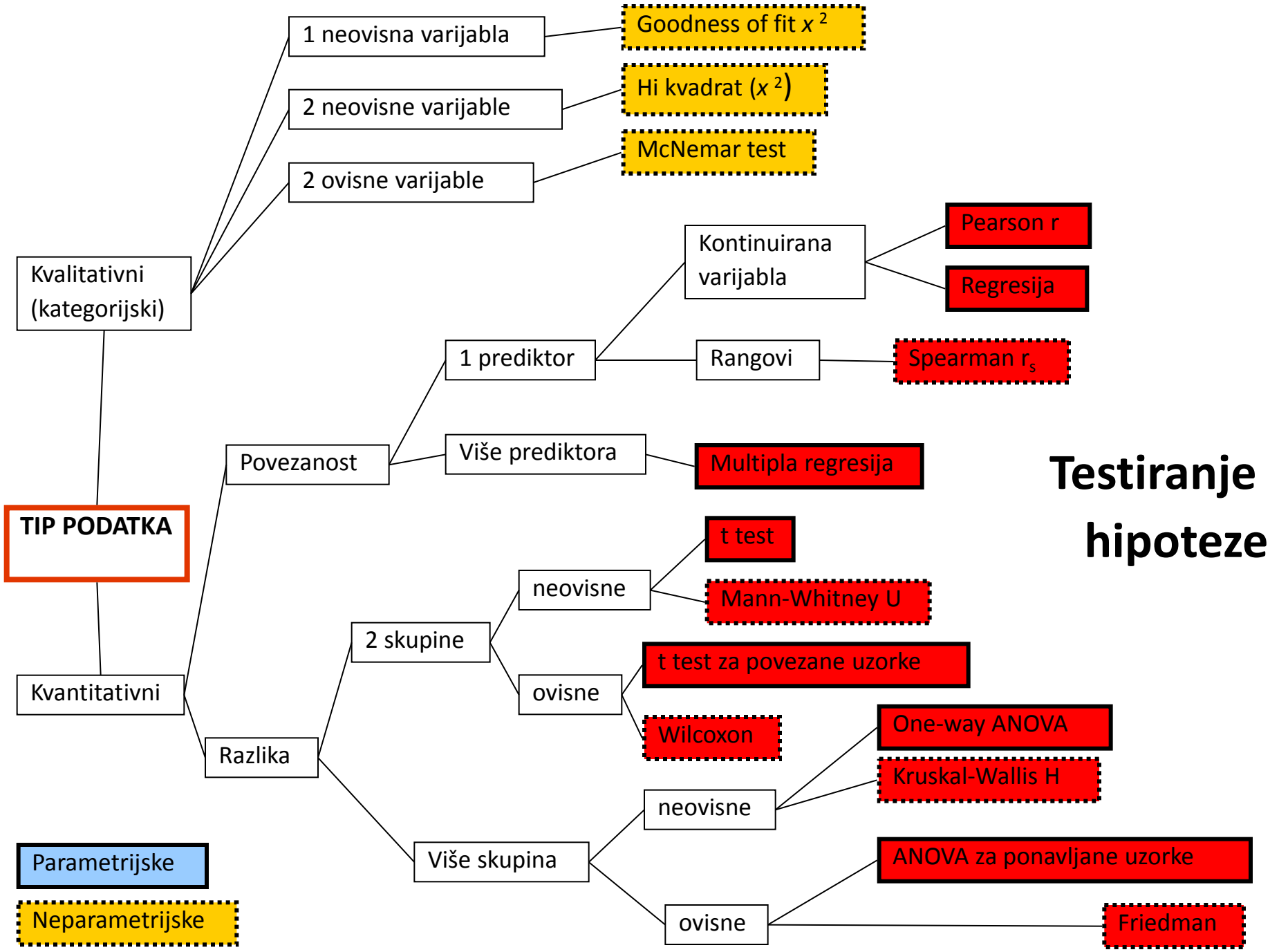


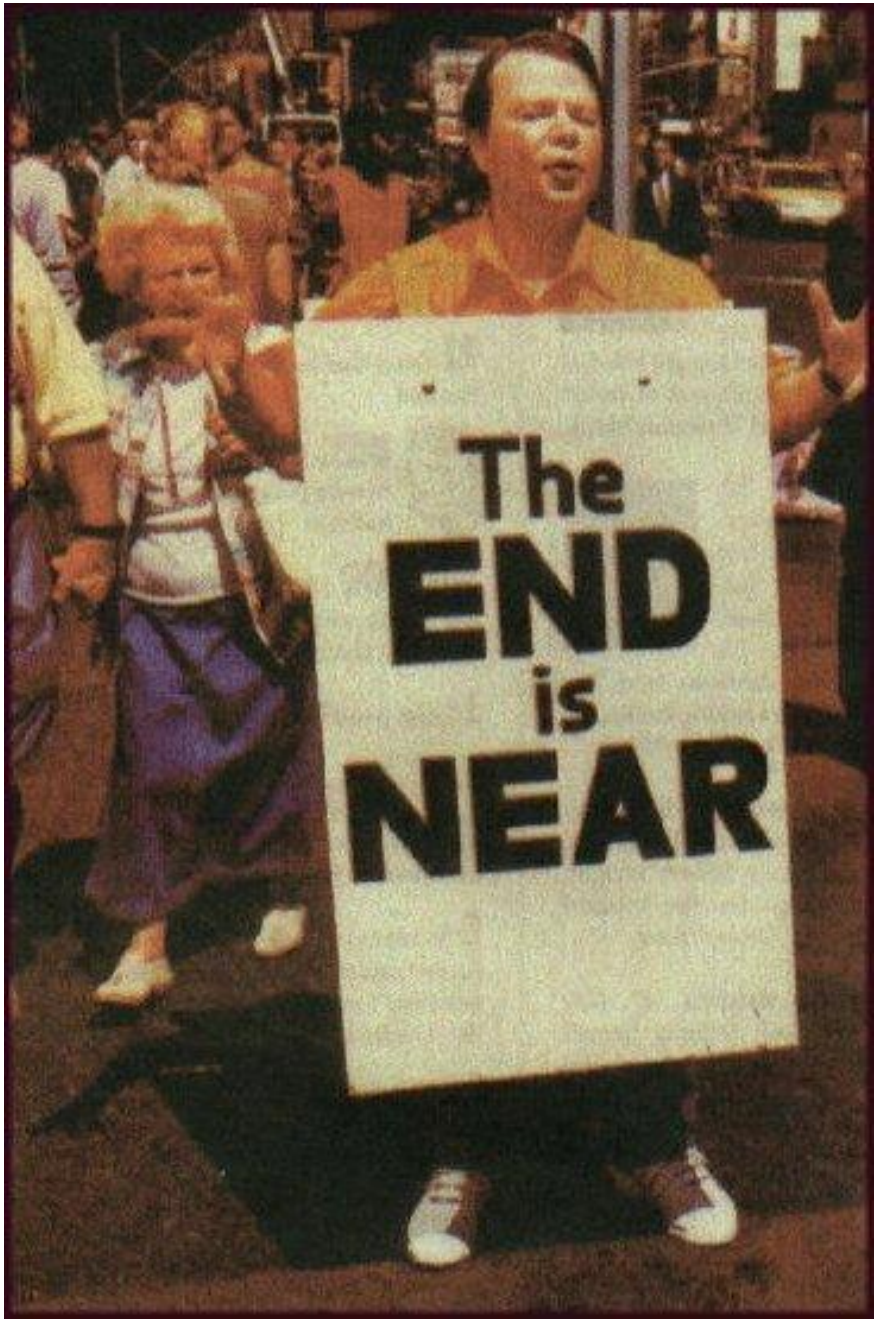
Poslije

Sistolički krvni tlak






Testiranje hipoteze
TIP PODATKA
Kvalitativni (kategorijski)
1 neovisna varijabla
Goodness of fit x^2
2 neovisne varijable
Hi kvadrat (x^2)
2 ovisne varijable
McNemar test
Kvantitativni
Povezanost
1 prediktor
Kontinuirana varijabla
Pearson r
Regresija
Rangovi
Spearman r_s
Više prediktora
Multipla regresija
Razlika
2 skupine
neovisne
t test
Mann-Whitney U
ovisne
t test za povezane uzorke
Wilcoxon
Više skupina
neovisne
One-way ANOVA
Kruskal-Wallis H
ovisne
ANOVA za ponavljane uzorke
Friedman
Parametrijske
Neparametrijske


The
END
is
NEAR

Hi-kvadrat test

- Jedan od najjednostavnijih statističkih testova
- Jako često se koristi
- Veliki broj neparametrijskih testova svodi se na hi-kvadrat

Hi-kvadrat - pažnja

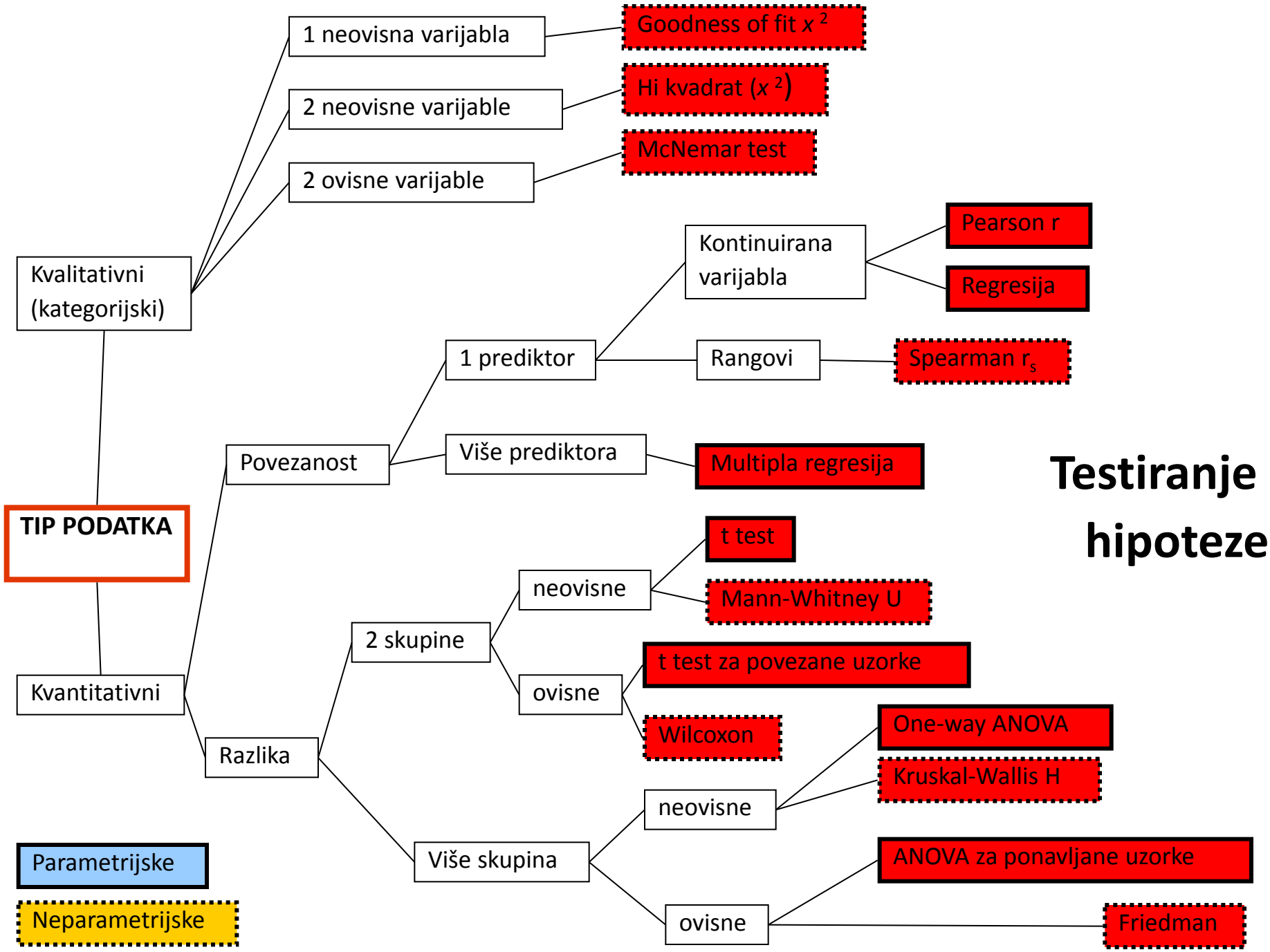
- Primjenjiv samo na kategorijskim podacima
- Primjeri:
  - Ocjene
  - Stupanj fizičke aktivnosti
  - Boja očiju
  - Spolne razlike
  - Socioekonomski status

Hi-kvadrat

- Temelji se na usporedbi očekivanih i opaženih frekvencija
- Za mali broj uzoraka (manji od 5 u 20% ili više polja tablice kontingencije) potrebno je koristiti Fisherov test

McNemar

- Alternativa hi-kvadrata za povezane varijable
- Kategorijske varijable koje su povezane
  - Pripadnost političkoj stranci prije i nakon izbora
  - Ishod liječenja u *cross-over* pokusu


Testiranje hipoteze
TIP PODATKA
Kvalitativni (kategorijski)
1 neovisna varijabla
Goodness of fit x^2
2 neovisne varijable
Hi kvadrat (x^2)
2 ovisne varijable
McNemar test
Kvantitativni
Povezanost
1 prediktor
Kontinuirana varijabla
Pearson r
Regresija
Rangovi
Spearman r_s
Više prediktora
Multipla regresija
Razlika
2 skupine
neovisne
t test
Mann-Whitney U
ovisne
t test za povezane uzorke
Wilcoxon
Više skupina
neovisne
One-way ANOVA
Kruskal-Wallis H
ovisne
ANOVA za ponavljane uzorke
Friedman
Parametrijske
Neparametrijske

Analitička statistika - testiranje hipoteze

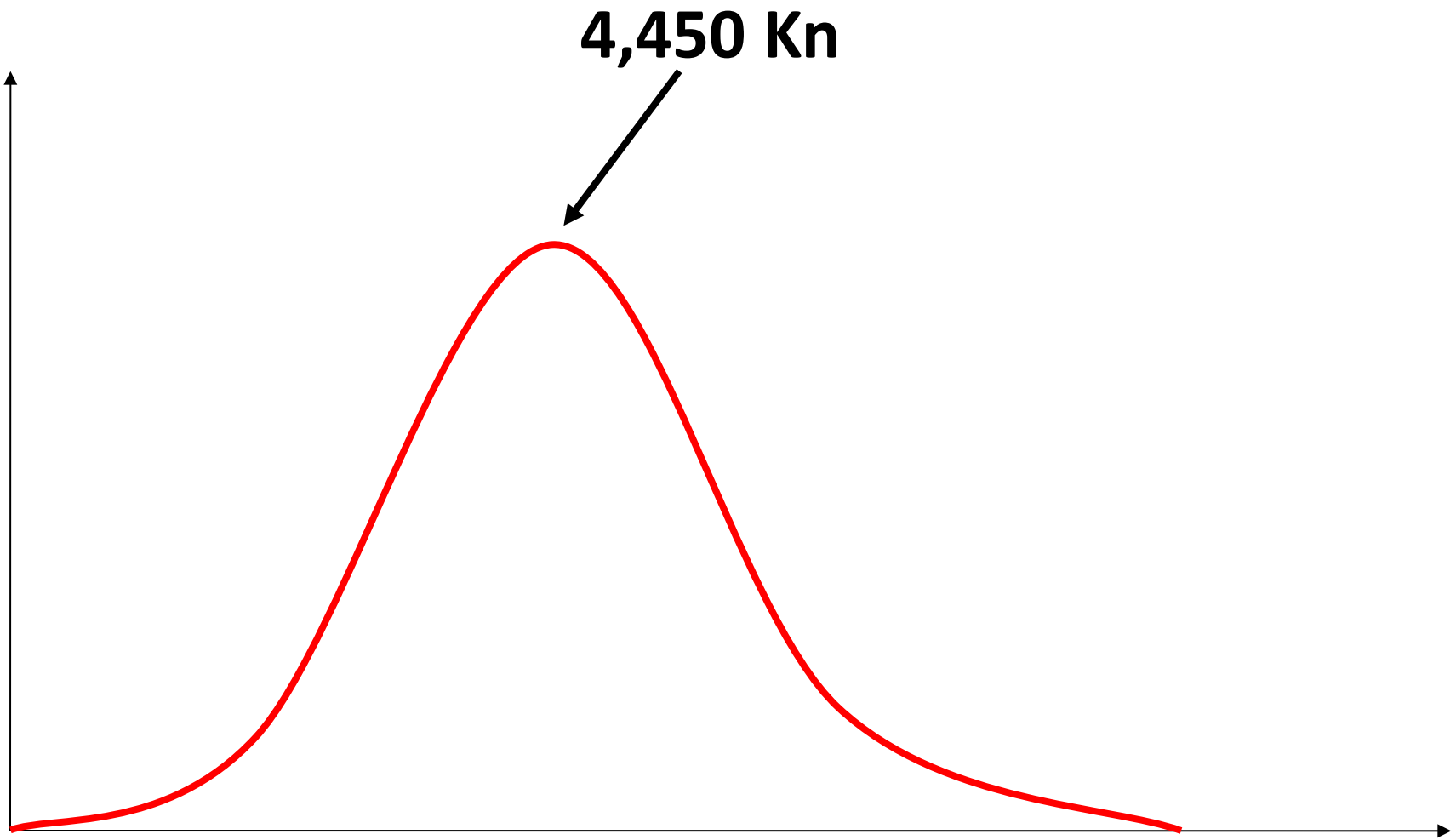
- P vrijednost (eng. *probability* - vjerojatnost)
- Govori o tome kolika je vjerojatnost da je rezultat točan, tj. da nije točan
- Manja od 0,05 (ili 0,01) – govori o tome da je vjerojatnost slučajnog i netočnog rezultata manja od 5% (1%)

Analitička statistika - testiranje hipoteze

- P vrijednost
- Manja od 0,05 (ili 0,01)
  - $P<0,05$
  - P NS.
  - $P=0,021$
  - $P<0,001$
  - $P=3,45*10^{-5}$
- Odabir statističkog testa ovisno o istraživačkom pitanju, obilježjima analiziranih varijabli i strukturi istraživanja

Primjer (1/2)

- Prosječna plaća u Republici Hrvatskoj iznosi 4.450 Kn
- U gradu Zagrebu 5.097 Kn


4,450 Kn

Najveća plaća u Hrvatskoj 741.824 kune

Business.hr

06.07.2007 09:12
Pročitaj komentare (0)
Ocijenite članak (0)

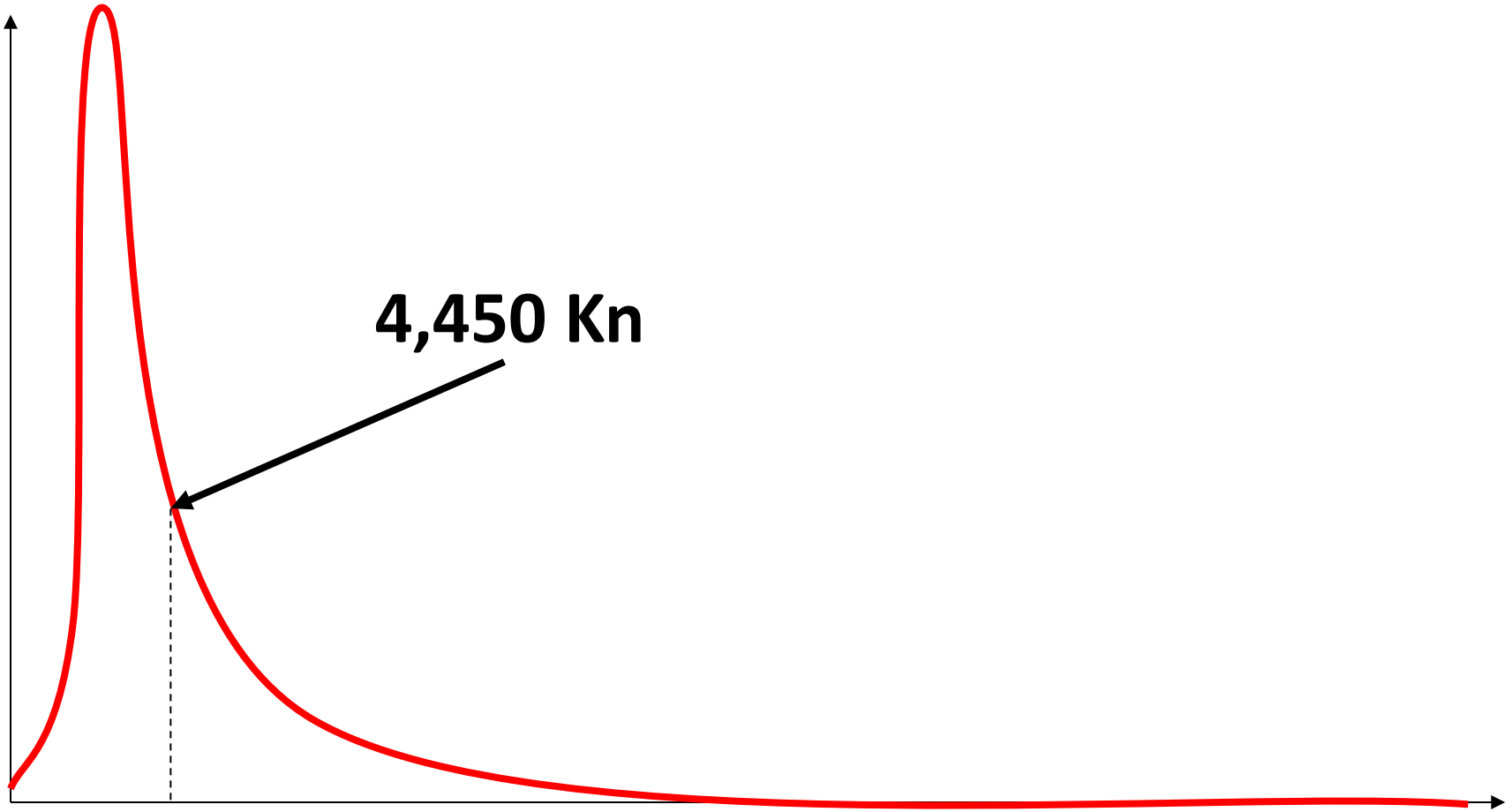
Prema do sada obrađenim podacima Porezne uprave, najveća plaća u Hrvatskoj lani je iznosila 741.824 kune, a isplaćena je jednom Varaždincu koji je tako godinu završio bogatiji za 8,9 milijuna kuna, piše Slobodna Dalmacija.

Snimio Hrvoje Knez

Drugi na listi 20 najvećih prošlogodišnjih dohodaka je Zagrepčanin čija godišnja plaća iznosi 8,8 milijuna kuna, a zanimljivo je da su, osim prvoplasiranog, svi na listi iz Zagreba.

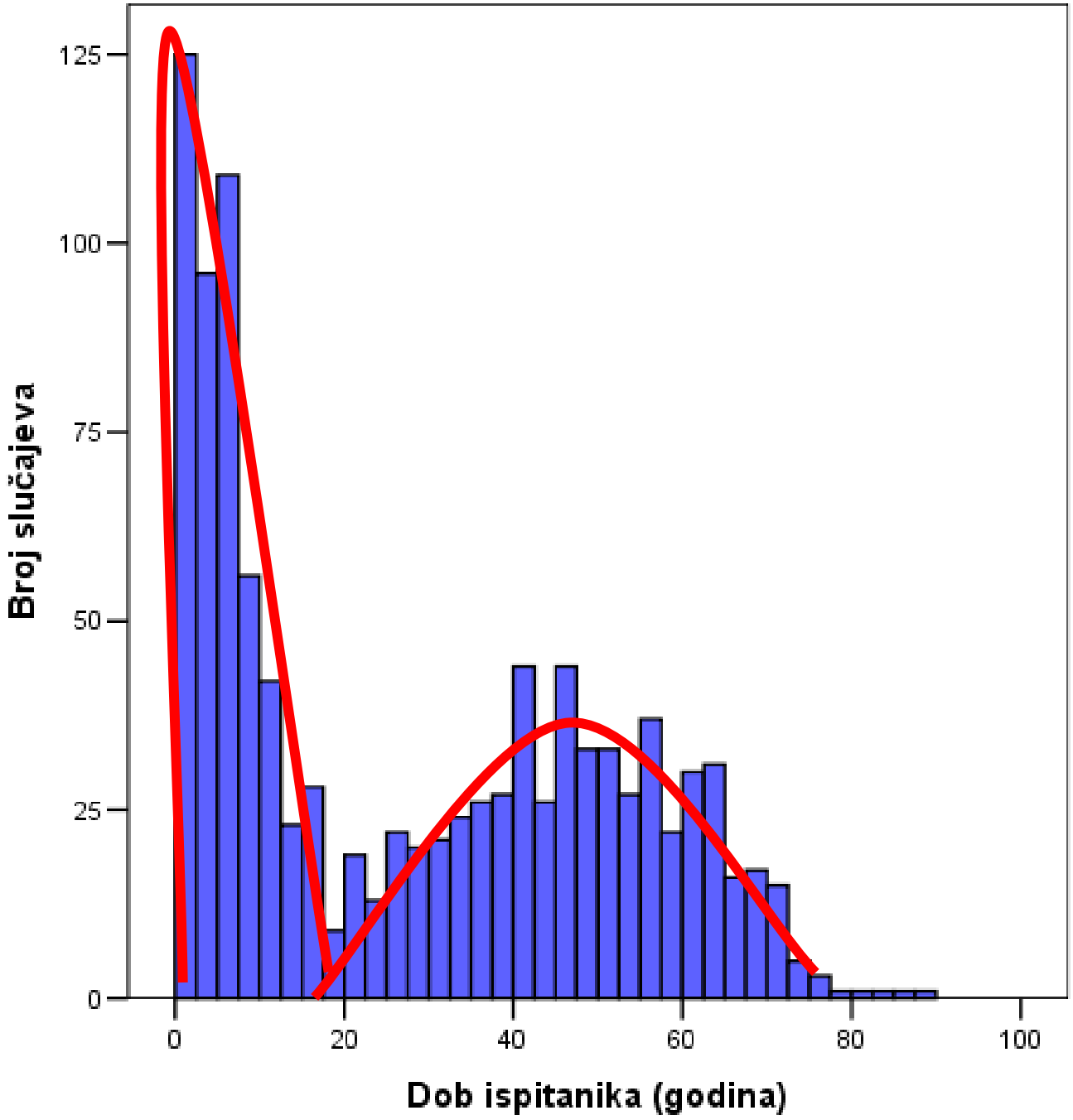
Najveći prijavljeni dohodak među obrtnicima iznosi 3,07 milijuna kuna, a ostvario ga je Osječanin koji se bavi visokogradnjom. Drugi obrtnik po zaradi je također Osječanin - arhitekt koji je zaradio 2,04 milijuna kuna. Na listi 20 najvećih zarada od obrta i slobodnih zanimanja najviše uvrštenih - njih osam - bavi se građevinarstvom te iznajmljivanjem i kupoprodajom nekretnina, a četvero su odvjetnici.

Novine


4,450 Kn

Primjer (2/2)

- U razdoblju od 1995-2005 godine u Vinkovcima je zabilježeno 1102 slučaja alergijskog rinitisa i astme. Srednja dob svih ispitanika bila je 24,3±11,6 godina. Odnos spolova bio je podjednak, 50,3% uzorka bili su muškarci.


125
100
75
50
25
0
Broj slučajeva
0
20
40
60
80
100
Dob ispitanika (godina)